Variation in the *Anopheles gambiae TEP1* Gene Shapes Local Population Structures of Malaria Mosquitoes

D i s s e r t a t i o n

Zur Erlangung des akademischen Grades

D o c t o r r e r u m n a t u r a l i u m

(Dr. rer. nat.)

Im Fach Biologie

eingereicht an der

Lebenswissenschaftlichen Fakultät

der Humboldt-Universität zu Berlin

von

BSc. (Biochemistry and Molecular Biology), MSc. (Biochemistry)

Evans Kiplangat Rono

Präsidentin der Humboldt-Universität zu Berlin:

Prof. Dr.-Ing. Dr. Sabine Kunst

Dekan der Lebenswissenschaftlichen Fakultät:

Prof. Dr. Bernhard Grimm

Gutachter/innen:

1. Dr. Elena A. Levashina
2. Prof. Dr. Arturo Zychlinski
3. Prof. Dr. Susanne Hartmann

Eingereicht am: Donnerstag, 04.05.2017

Tag der mündlichen Prüfung: Donnerstag, 29.06.2017

Zusammenfassung

Rund eine halbe Million Menschen sterben jährlich im subsaharischen Afrika an Malaria Infektionen, die von der *Anopheles gambiae* Mücke übertragen werden. Die Allele (**R1*, **R2*, **S1* und **S2*) des *A. gambiae* complement-like thioester-containing Protein 1 (TEP1) bestimmen die Fitness der Mücken, welches die männlichen Fertilität und den Resistenzgrad der Mücke gegen Pathogene wie Bakterien und Malaria-Parasiten. Dieser Kompromiss zwischen Reproduktion und Immunnität hat Auswirkungen auf die Größe der Mückenpopulationen und die Rate der Malariaübertragung, wodurch der *TEP1* Lokus ein Primärziel für neue Malariakontrollstrategien darstellt. Wie die genetische Diversität von *TEP1* die genetische Struktur natürlicher Vektorpopulationen beeinflusst, ist noch unklar. Die Zielsetzung dieser Doktorarbeit waren: i) die biogeographische Kartographierung der *TEP1* Allele und Genotypen in lokalen Malariavektorpopulationen in Mali, Burkina Faso, Kamerun, und Kenia, und ii) die Bemessung des Einflusses von *TEP1* Polymorphismen auf die Entwicklung humaner *P. falciparum* Parasiten in der Mücke. Informative Einzelnukleotid-Polymorphismen (SNPs) im *TEP1* Lokus wurden identifiziert und als genetischer Marker für PCR-Restriktions-Fragment-Längenpolymorphismus (PCR-RFLP) Hochdurchsatz-Genotypisierung von im Feld gefangenen Mücken-Proben validiert. Wir haben ein neues Allel identifziert, hier als **R3* benannt, welches ausschließlich in *A. merus* Populationen in Kenia existiert. Die Verteilung von *TEP1* Allelen und Genotypen in Populationen verschiedener Mückenarten wurden in spezifische biogegraphische Gruppen in vier ausgewählten Ländern im subsaharischen Afrika kategorisiert. Die Analysen der *TEP1* Polymorphismen zeigten, dass die natürliche Selektion auf Exone, sowie Introne wirkt, was auf eine starke funktionale Beschränkung an diesem Lokus hindeutet. Außerdem zeigen unsere Daten die strukturierte Erhaltung natürlicher genetischer Variation im *TEP1* Lokus, in welchem die Allele und Genotypen spezifische evolutionäre Wege verfolgen. Diese Ergebnisse weisen auf die Existenz von arten- und habitatspezifischen Selektionsdrücken hin, die auf den *TEP1* Lokus wirken. Des Weiteren habe ich den Einfluss der *TEP1* Polymorphismen auf die Mückenresistenz gegen *P. falciparum* in experimentellen Infektionen evaluiert. Meine Resultate haben gezeigt, dass *TEP1*S1* und **S2* Mücken gleichermassen empfänglich für *Plasmodium*-Infektionen sind. Außerdem habe ich einen hohe Sterblichkeitsrate in der **R1/R1* Laborkolonie im Vergleich zu den empfänglichen Mücken-Linien beobachtet. Da die **R1/R1* Mücken

ausschließlich in der *A. coluzzii* Mückenart in Westafrika gefunden wurden, postuliere ich, dass **R1* ein konditional-lethales Allel darstellt, welches gewisse, noch unbekannte Bedingungen für erfolgreiche Erhaltung und Verbreitung benötigt. Insgesamt tragen die Resultate der biogeographischen Kartographierung des *TEP1* Lokus und der Züchtungs- und Infektionsexperimente zu einem besseren Verständnis über den Einfluss der verschiedenen Vektorarten und lokale Umwelteinflüsse auf die Vektorpopulationen und Malariaübertragung bei. Des weiteren kann die hier beschriebene hochdurchsatz-genotypisierungs Methode, zur Studie lokaler *A. gambiae* Mückenpopulationen, in der Feldforschungsarbeit eingesetzt werden. Dieser neue Ansatz wird die epidemiologisch relevante Überwachung und Vorhersage dynamischer Prozesse in lokalen Malariavektorpopulationen unterstützen, welche die Entwicklung neuer Strategien der Vektorkontrolle ermöglichen könnten.

Abstract

About half a million people die annually in sub-Saharan Africa due to malaria infections transmitted by *Anopheles gambiae* mosquitoes. The alleles (**R1*, **R2*, **S1* and **S2*) and the genotypes of *A. gambiae* complement-like thioester-containing protein 1 (TEP1) determine the fitness in male fertility and the degree of mosquito resistance to pathogens such as bacteria and malaria parasites. Because this trade-off between the reproduction and the immunity impacts directly on mosquito population abundance and malaria transmission, respectively, the *TEP1* locus is a prime target for new malaria control strategies. How *TEP1* genetic diversity influences the genetic structure of natural vector populations and development of human malaria parasites is unclear. The aims of this thesis were to: i) map distribution of *TEP1* alleles and genotypes in local malaria vector populations in Mali, Burkina Faso, Cameroon and Kenya, and ii) assess the impact of *TEP1* polymorphism on development of human *P. falciparum* parasites in mosquitoes. Informative single nucleotide polymorphisms (SNPs) at *TEP1* locus were identified and validated as genetic markers for PCR-restriction fragment length polymorphism (PCR-RFLP) high-throughput genotyping of the field mosquito samples. We identified a new allele, herein named **R3,* that is private to *A. merus* populations in Kenya. The distribution of *TEP1* alleles and genotypes in mosquito species populations were categorized into specific biogeographic groups across the four selected countries in sub-Saharan Africa. Analyses of *TEP1* polymorphism revealed that natural selection acts in concert on both exons and introns, suggesting strong functional constrains acting at this locus. Moreover, our data demonstrate a structured maintenance of natural *TEP1* genetic variation, where the alleles and the genotypes follow distinct evolutionary paths. These findings suggest the existence of species- and habitat-specific selection patterns that act on *TEP1* locus. I further evaluated contribution of *TEP1* polymorphism on mosquito resistance to *P. falciparum* in experimental infections. My results revealed that the *TEP1*S1* and **S2* mosquitoes are equally susceptible to *Plasmodium* infections. I also observed high mortality rates of the **R1/R1* laboratory colony as compared to the susceptible lines. As the **R1/R1* mosquitoes were restricted to the *A. coluzzii* species and to the specific ecology in West Africa, I propose that **R1* is a conditional lethal allele, which requires certain yet unknown conditions for successful breeding and maintenance. Collectively, results of my thesis on the biogeographic *TEP1* mapping, and on the breeding and infection experiments contribute to a better understanding of

how the vector species and local environmental factors, shape vector population structures and malaria transmission. Furthermore, the high throughput *TEP1* genotyping approach reported here could be used for field studies of local *A. gambiae* mosquito populations. This new approach will benefit surveilance and prediction of dynamics in local malaria vector populations that may have epidemiological significance, and therefore inform the development of novel vector control measures.

Table of Contents

List of Figures

List of Tables

Dedication

To Sarah, my mother for taking me to school and inspiring my dreams to reality.

To Mercy, my dear friend and lovely wife for selflessly supporting me.

To Victor and Emmanuel, my great sons for bringing joy into my life.

Declarations

1. Collaborations

Below, I acknowledge and provide the details of the scientific collaborations without which this study would not have been successful.

In chapter 2, the study was the European Commission FP7 Cooperative Project "MALVECBLOK" (CNRS, France; UNIPG, Italy; RUNMC, The Netherlands; IRD, France; UNIROMA, Italy; MRTC, Mali; ICIPE, Kenya). Elena A. Levashina (EAL) was the Project Coordinator. The contributors were in Mali [Alou G. Traoré (AGT), Modibo Mariko, Djibril Sangare (DS), Mouctar Diallo (MD)]; Burkina Faso [Julien Pompon (JP), Francesco Baldini (FB), Yannis Thailayil (YT), Priscila Bascunan (PB), Flaminia Catteruccia (FC), Roch Dabire (RD) and Abdoulae Diabate (AD)]; Cameroon [Isabelle Morlais (IM), Anne Boisiere (AB), Sandrine Nsango (SN) and Parfait Awono-Amebe (PAA)]; Kenya [Daniel Masiga (DM), Paul Mireji (PM), Pamela B. Seda (PBS), Martin K. Rono (MKR) and me]; and Germany [Hanne Krüger (HK), Markus Gildenhard (MG), EAL and myself]. Specifically, FC, IM and EAL conceived and designed the study. DS, MD, AD, RD, SN, PAA, DM, IM managed the sample collections, genotyping and sequencing. AB, IM, EAL and I, contributed the molecular tools. YT, FB, JP, PAA, AB, PBS, PM, AGT, DS, MD, MKR, HK and I, conducted sample collections, processing and genotyping. PB, AB, PM, PBS and I, performed sequencing. MG, EAL and I, contributed ideas during the data analyses. I analyzed the data presented in the thesis and composed the thesis.

In chapter 3, the experiments were conceived and designed by EAL and me. I conducted the experiments and the data analyses.

This work has not been submitted elsewhere for the purpose of obtaining a degree or professional qualification.

Evans Kiplangat Rono

Berlin, 03.05.2017

2. Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt, die vorliegende Dissertation selbstständig angefertigt und keine anderen als die angegebenen Hilfsmittel verwendet zu haben.
Ich erkläre hiermit, dass ich an keiner anderen Universität ein Prüfungsverfahren beantragt bzw. die Dissertation in dieser oder anderer Form bereits anderweitig als Prüfungsarbeit verwendet oder einer anderen Fakultät als Dissertation vorgelegt habe.
Wurden Ergebnisse in Kooperation produziert, ist dies entsprechend angegeben.
Die vorliegende Arbeit wurde am Max Planck Institut für Infektionsbiologie unter Leitung von Prof. Dr. Elena Levashina durchgeführt.

Evans Kiplangat Rono

Berlin, 03.05.2017

3. Erweiterte Eigenständigkeitserklärung

Hiermit versichere ich, Evans Kiplangat Rono, dass folgende Publikation:
"Mosquito microevolution drives *Plasmodium falciparum* dynamics" (Arbeitstitel, Manuskript in Bearbeitung)
maßgeblich von mir verfasst wurden. Mögliche Übereinstimmungen mit Textpassagen aus meiner Dissertation " Variation in the *Anopheles gambiae TEP1* Gene Shapes Local Population Structures of Malaria Mosquitoes " stellen somit keinen Plagiatsfall dar.
Dies wird bei Bedarf durch die Betreuerin der Dissertation und Co-Autorin der aufgeführten Publikation Prof. Dr. Elena Levashina bestätigt.

Evans Kiplangat Rono

Berlin, 03.05.2017

Acknowledgements

First, I thank my Almighty God for love, grace, and blessing. He has been the fountain of my hope, happiness and strength.

I am thankful to my supervisor Dr. Elena A. Levashina and Prof. Dr. Arturo Zychlinsky, and the Max Planck Institute for Infection Biology and Humboldt University for hosting me and supporting my training. Particularly, Elena has been such a great intellectual advisor and a mentor to me. How lucky am I to have been shaped by your kindness, understanding and patience, and your training to grow bigger in Science! *Спасибо.*

I appreciate the members of my thesis committee Prof. Dr. Kai Matuschewski, Dr. Elena A. Levashina, Prof. Dr. Arturo Zychlinski, Prof. Dr. Susanne Hartmann and Prof. Dr. Thomas F. Meyer. Many thanks go to my thesis reviewers; Dr. Elena A. Levashina, Prof. Dr. Arturo Zychlinski and Prof. Dr. Susanne Hartmann for reviewing this work.

I thank Dr. Daniel Masiga and Mr. James Kabii of *icipe,* Kenya for mentoring and inspiring me.

I am grateful to the DAAD for awarding me a PhD fellowship.

I thank the European Commission FP7, EVIMalar and MALVECBLOK for funding the MALVECBLOK project.

I appreciate all the MALVECBLOK collaborators and contributors in Mali, Cameroon, Burkina Faso and Kenya for their tremendous contribution to this work.

I acknowledge Dr. Joseph Mwangangi and Mr. James Kabii for the sampling facilitation in Kenya. I thank Mr. Charles Okech for helping in sampling in Ahero, Kenya. Special thanks go to Mr. David Shida, Mr. Gabriel Nzai and Mr. Peter Njue for their assistance in sampling in Malindi, Kenya.

I am particularly grateful to Mr. Collins Omogo and Miss. Josphine Shikaya for assisting during PCR optimization and genotyping. Many thanks to Mrs. Judy Mwaura and Dr. Mercy Mwaniki for their assistance in the analysis of GIS data and designing of geographical maps.

I acknowlegdge the following collaborators for kindly providing us with mosquito and parasite strains: Dr. Flaminia Cateruccia (G3 mosquitoes), Paul Howell of MR4 (4Arr mosquitoes), Dr. Stéphanie Blandin and Dr. Eric Marois (L3-5 and Mut6 mosquitoes), and Prof. Robert Sauerwein (NF165 and NF54HT-GFP-luc *P. falciparum* parasites).

This work would not have been successful without the organization skills from our gifted laboratory manager, Dr. Yara Reis. Besides, she was always concern and caring to me, more so on social life outside the laboratory especially the wellbeing of my family and kids.

I salute Dr. Giulia Costa for her great input and suggestions on our parasite cultures and mosquito breeding strategies. I recognize my supervisor Dr. Elena Levashina and Markus Gildenhard for our fruitful discussions and writing of the manuscript from the data in Chapter 2 of this thesis.

In no special order, I appreciate the help from our able technical crew Liane Spohr, Sandrina Koppitz, Hanne Krüger, Daniel Eyermann, Dana Tschierske, Cemil Yilmaz, Cynthia Yapto, Danja Sumpf, Jennifer Schmidt, Manuela Andres, and Maria Pissarev.

I am grateful for the help on genotyping spree from Valentina Rausch, Alberto Stella, and Cynthia Yapto. I am grateful and indebted to all the Vector Biology (current and the former) group mates and the department of Parasitology for the company, quality interactions and fruitful discussions. To Dr. Lena Lampe, Dr. Ewa Chrostek and Dr. Philip Hügli, *Danke schön* you guys for being brilliant, nice and supportive office mates. To Christine Kappler, it was always refreshing to have your French Gheese whenever you came from Strasbourg to visit our laboratory in Berlin.

I thank my current and past DAAD-Kenyan community in Berlin, church friends in BICC and IDC among others for social network and interaction.

Thank you Dr. Joel L. Bargul and Elias Mibei for being my special friends, *asanteni sana ndugu zangu.*

I thank my mother (Sarah), brothers, sisters, nephews and nieces among others for your support and encouragement.

Importantly, I am deeply indebted to my dear wife (Mercy) and my sons (Victor and Emmanuel) for the unwavering love, prayers and support in times of highs and lows. *Kongoi missing.*

Acronyms and Abbreviations

%	percentage
°C	Degree Celcius
µ	micro
µg	microgram
µl	microliter
µM	micromolar
AP	Alternative Pathway
bp/kb	Base pair/ Kilobase
cm	centimeter
CP	Classical Pathway
DMSO	Dimethyl sulfoxide
DNA	Complementary deoxyribonucleic acid /deoxyribonucleic acid
dNTPs	Deoxyribonucleic acid
dsRNA	double stranded messenger RNA
DTT	Dithiothreitol
e.g.	for example
EDTA	Ethyle diamine tetra-acetic acid
et al.	*Et alia* (and the colleagues)
FACS	Fluorescence activated cell sorting
gDNA	genomic DNA
GFP	green fluorescence protein
GFP	Green fluorescent protein
h	hour (s)
i.e.	that is

ICIPE	International Center of Insect Physiology and Ecology
IMBC	Institut de Biologie Moléculaire et Cellulaire, France
IMD	Immune Deficiency Pathway
JAK/STAT	Janus Kinase /Signal Transducer and Activation of Transcription
KDa	kilodalton
l	liter
LP	Lectin Pathway
M	Molar
M/mM	Molar/Millimolar
min (s)	minute(s)
mL/µl	Millilitres/Microlitres
mM	Millimolar
mRNA	Messenger RNA
NCBI	National Center for Biotechnology Information
NFW	Nuclease free water
NIAID	National Institute of Allergy and Infectious Diseases
ORF	Open reading frame
PBS	Phosphate buffered saline
pmol	picomole
pmol/µl	Picomoles per microlitres
RBC	red blood cell
RNA	Ribonucleic acid
RNAi	Ribonucleic Acid interference
RT	Room temperature
s	seconds

spp.	Species (in plural)
T.A.E	Tris-acetate-EDTA
T.E	Tris-EDTA
Taq	*Thermus aquaticus*
U	enzyme unit
UV	Ultra-violet
V	volts
WHO	World Heath Organization

Chapter 1

General Introduction

General Introduction

1.1 Summary

The female mosquito *Anopheles gambiae* species transmits deadly human malaria parasites. These parasites are protozoan pathogens belonging to the genus *Plasmodium. P. falciparum* is responsible for the deadliest cases of malaria leading to deaths in sub-Saharan Africa (1-3). *Plasmodium* life cycle takes place in both the mosquito vector and human hosts. In *A. gambiae*, complement-like TEP1 (Thioester-containing Protein 1) plays a significant role in elimination of malaria parasites (4, 5). TEP1 is encoded by an exceptionally polymorphic gene (6) whose allelic variation correlates with the distinct phenotypes in resistance to *Plasmodium* infections as well as in male fertility (5). However, little is known about mosquito ecological factors that shape natural genetic variability at the *TEP1* locus in mosquito populations along the geographic clines of sub-Saharan Africa (6-8). Moreover, how this genetic variation affects the development of *P. falciparum,* human malaria parasites, is poorly understood (6-8). In this context, high throughput genotyping strategies are needed in order to define local adaptation of malaria vector populations to different ecological niches. This chapter reviews our current understanding on the key dynamics of malaria transmissions and advances in understanding of the mosquito immune responses. It identifies research gaps, and introduces the aims and outline of this thesis.

1.2 Malaria in the world

Malaria is one of the deadliest human infectious diseases worldwide, and especially in sub-Saharan Africa where about 400,000 deaths occur annually (1-3). It is a mosquito-borne disease caused by the protozoan pathogen of the *Plasmodium* genus. The species of human malaria parasites include *P. falciparum, P. vivax, P. malariae, P. ovale* and *P. knowlesi* (2, 9, 10). However, *P. falciparum* infections contribute to the greatest burden for the most devastating deaths in sub-Saharan Africa (2, 9, 10). For example, in malarious holoendemic (i.e. where malaria transmission occurs throughout the year) regions, such as western Kenya, *P. falciparum* causes up to 100% of all the malaria cases accounting for 50% of all clinically diagnosed infectious diseases (11-13).

P. vivax is able to survive in cooler and higher altitudes, develop in the mosquito vector at lower temperatures, and undergo long dormant liver stages (2). However, in most African human populations, the presentation of malaria complications resulting

from *P. vivax* infection is relatively milder than those of *P. falciparum* malaria because many African human populations lack Duffy antigen (i.e. Duffy blood group antigen) on the surface of red blood cells (*1*). The Duffy antigen acts as a receptor for the invasion of red blood cells by the *P. vivax* and *P. knowlesi* malaria parasites (*14*). Human malaria cases caused by the *P. knowlesi* in South-East Asia are associated with zoonotic transmission, where mosquitoes acquire blood from an infected monkey and then pass the infection to humans during the subsequent blood feeding (*2*).

Mosquitoes are two-winged little flies of the *Culicidae* family, consisting of over 3500 species described so far (*1-3*, *15*). Human malaria vectors (carriers) are those mosquitoes that are infected with human *Plasmodium* malaria parasites and pierce human skin to acquire blood meal, and in the process, transmit the parasites to humans (*1-3*). Female *Anopheles* mosquitoes are the major and the most effective human malaria vectors that constitute about 30 of over 400 species of the *Anopheles* mosquitoes (*1*, *2*).

Globally, thanks to the deliberate efforts comprising vector control, chemoprevention and case management strategies to eliminate malaria, there is a significant drop of >35% in malaria cases and deaths in the last 15 years (*1*, *2*, *16*). However, despite this drop, incidences of global malaria infections are still high. For instance in 2015 alone, the total malaria cases and deaths worldwide were 214 million and 438,000, respectively (*2*). In addition, most malaria cases (88%) and most deaths (90%) occurred in the malaria endemic regions of Africa (*2*). Moreover, the currently existing vector control strategies are limited to the use of insecticide treated bed nets (ITNs) and indoor residual spraying (IRS), which are suffering a severe drawback due to resistance development in mosquito vector (*1*).

The increasing knowledge on the mosquito's immune responses against pathogens such as *Plasmodium* parasites could provide promising alternatives towards the development of novel vector-based malaria control strategies (*5*, *17*, *18*).

1.3 Malaria transmission cycle

1.3.1 Life cycle of the malaria mosquito

The life cycle of the mosquito is relatively short, involving egg, larvae, pupae and adult stages i.e. complete metamorphosis (**Fig. 1-1**). The adults feed on sugars from nectar and in particular, the females require blood meal as source of protein and lipids for egg development (*15*). Thus, the females have to actively look for a vertebrate host

to bite and acquire the blood meal from. It takes about 48 h post-blood-meal acquisition and mating, for the eggs to develop. The female then searches for stagnant water and oviposits about 200 eggs onto the surface. The egg hatching, larvae, pupae stages are aquatic and the period between the egg and the pupae takes 1 to 2 weeks depending on the species, food and temperature conditions. The larval stage consists of four instars (phases) where the larvae molts four times, as it grows larger and larger before ultimately reaching the pupae stage. The terrestrial life of the adult stage can take up to a month on average (*15*).

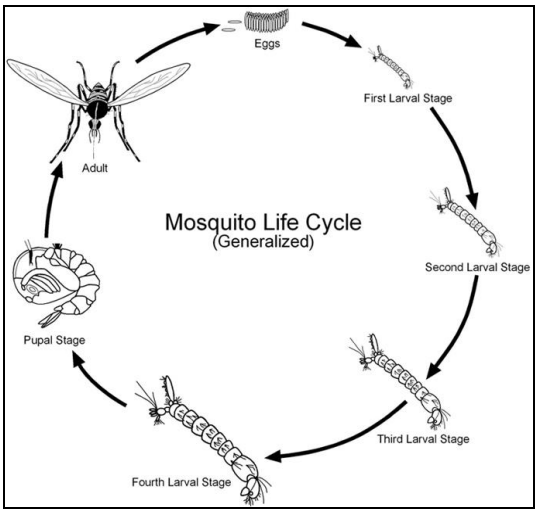


Fig. 1-1. General life cycle of the mosquito.
The life cycle of a mosquito undergoes complete metamorphosis i.e. egg, larvae, pupae and adult stages. The stages of eggs, larvae and pupae are aquatic while the adult stage is terrestrial. In each stage, life span is dependent on the species, food and temperature conditions. For an adult female to lay eggs, it takes up a blood meal from a vertebrate host, and gets mated by a male in the male mosquito swarm. Two days later, the female lays about 200 eggs on the surface of water. The eggs hatch to larvae after 24 h. The larval phases consist of 4 stages i.e. instars in between which molting occurs as the larvae grows bigger. Pupae stage does not feed and lasts for about 24 h during which the adult body parts are formed. The adult emerges from the pupae and flies away to start the terrestrial life. Figure source: Scott Charlesworth, Purdue University.

1.3.2 Lif e cycle of the malaria parasite

The life cycle of *Plasmodium* parasite is very complex, comprising of multistage sexual and asexual developmental stages that take place in two hosts; a primary host (female *Anopheles* mosquito vector) and a mammalian or vertebrate host (secondary host) (*9*, *10*) (**Fig. 1-2**). Both sexual and asexual replications occur in the mosquito vector, whereas asexual replications involving two distinct cycles in liver and in red blood cells, take place in a vertebrate host (**Fig. 1-2**).

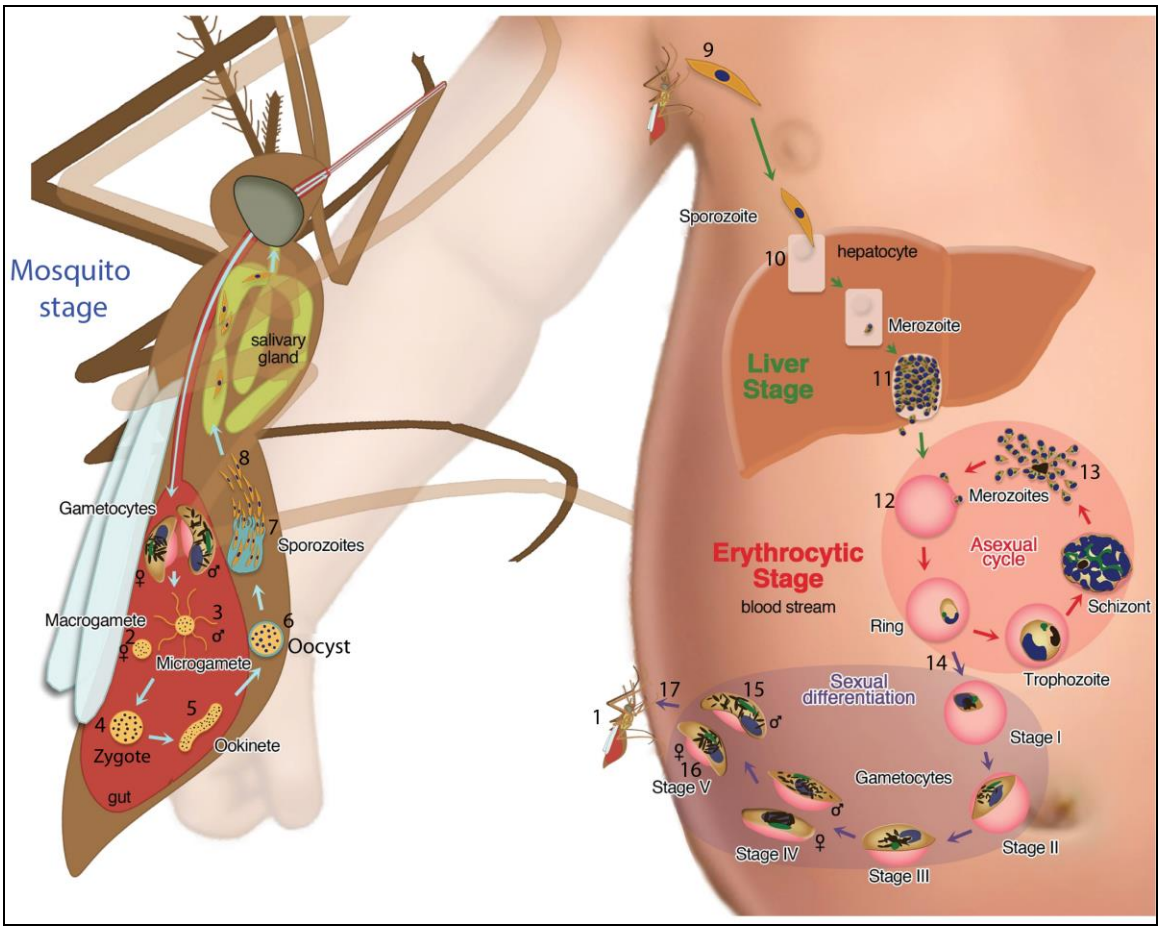


Fig. 1-2. Transmission cycle of human *Plasmodium* parasite.
A female mosquito (1) is infected with the parasite gametocytes as it takes a bloodmeal from an infected human host. The ingested female gametocyte differentiates into a single haploid female macrogamete (2), whereas the male gametocyte exflagellates and forms eight haploid male microgametes (3). The macrogamete and microgamete fuse and fertilization takes place to form a diploid zygote (4) which then develops into a motile ookinete (5). The ookinete invades the midgut epithelium of the mosquito and moves to the basal lamina to develop into an oocyst (6). The oocyst undergoes asexual replication cycles to produce haploid sporozoites (7). The sporozoites are released into the haemocoel when mature oocysts rapture. Sporozoites migrate to and invade the mosquito's salivary glands (8) and the infected mosquito is able to transmit (9) the parasite to the human host during feeding. The sporozoites injected into the human skin, reach the blood stream and invade the liver cells (10). In the liver cells or hepatocytes, the sporozoites undergo asexual replications to produce thousands of merozoites (11). The infected hepatocyte raptures to release mature merozoites that infect the red blood cells (RBCs) (12). Within the RBC, the merozoite undergoes rounds of asexual replications and transitions through a series of developmental stages resulting in thousands of merozoites (13) that are released to infect more RBCs upon the rapturing of the schizont. Some merozoites (14) differentiate into male (15) and female gametocytes (16), which circulate in the blood stream. The gametocytes (17) taken up by a mosquito during acquisition of the infected blood meal, complete the transmission cycle.
Figure source: Le Roch Laboratory, UC Riverside, adapted with modifications.

Mosquitoes are infected with *Plasmodium* parasites during the acquisition of blood meal from an infected mammalian host carrying male and female gamete-precursor cells called gametocytes (*10*, *19*, *20*). In the mosquito midgut, these gametocytes differentiate in about 15 min and mature into gametes through a process called

gametogenesis. The female gametocyte differentiates into a single haploid female macrogamete, whereas the male gametocyte undergoes exflagellation process, which gives rise to eight (8) haploid male microgametes. Within 1 h of acquisition of an infected blood meal, macrogametes and microgametes fuse and fertilization takes place to form a diploid zygote that develops within 18-20 h into a motile ookinete [reviewed in (*21*)]. The ookinetes migrate to cross the midgut epithelium of the mosquito and move to the basal lamina within 24 h post-infection [reviewed in (*21*)]. At this stage, the mosquito's immune system efficiently eliminates most *Plasmodium* parasites (*4*).

In about 2 days post-blood meal, ookinetes that survive the mosquito's defensive responses develop into oocysts, which undergo asexual replication cycles to produce haploid sporozoites (*22*). Subsequently, the mature oocysts rapture a week later to release midgut sporozoites into the haemocoel. Sporozoites migrate to and invade the salivary glands (*22*). At the salivary gland sporozoite stage, the infected mosquito is able to transmit the parasite to its vertebrate host during feeding.

Sporozoites that are injected into mammalian blood stream are invasive to the liver cells (hepatocytes) (*9*). Sporozoites in the liver stage undergo asexual replications for 6-10 d to produce thousands of merozoites, but at this stage the patient shows no clinical symptoms (*9*). In the liver, some *P. vivax* and *P. ovale* malaria parasites can remain dormant (i.e. undergo latency period) for months to years with occasional malaria relapses i.e. reoccurrence of malaria episodes due to the liver dormant malaria parasites without new infections (*10*).

Infected hepatocyte raptures to release mature merozoites that initiate the blood stage by infecting red blood cells (RBCs) (*9*). This period is also called an erythrocytic or blood schzogony stage. It involves asexual replication within the infected erythrocyte and undergoes three developmental stages (ring, trophozoite and schizont) producing dozens of merozoites per schizont. Schizont's rapture releases the merozoites that in turn infect more RBCs. Duration of the sporogonic process varies between *Plasmodium* species. For instance, it takes about 48 h in *P. vivax* and *P. ovale,* 72 h in *P. malariae* and 12 h in *P. knowlesi* (*23*).

Some merozoites differentiate into male and female gametocytes, which will not invade red blood cells but stay in blood circulation (*9*). Only the gametocyte stage is infecting the mosquito vector as it acquires the infected blood meal, and the transmission cycle begins all over again.

1.4 Distribution of species of malaria vectors in Africa

1.4.1 The sibling species of the *Anopheles gambiae* complex

The malaria vector species are distributed across the world, but the most efficient vectors are found in sub-Saharan Africa (3, 24) (**Fig. 1-3**; **Fig. 1-4**) (25). The *A. gambiae* complex, often referred to as *A. gambiae sensu lato* (*A. gambiae s.l.*), consists of the following eight cryptic (i.e. morphologically indistinguishable) African species: *A. gambiae sensu stricto* (*A. gambiae s.s.*), *A. arabiensis, A. coluzzii, A. merus, A. melas, A. quadriannulatus, A. amharicus* and *A. bwambae* (25-29). The following five species were first described through laboratory crosses that resulted in hybrid sterility (26): Three fresh water species; *A. gambiae s.s.*, *A. arabiensis* and *A. quandriannulatus,* and two salty water species; *A. melas* in West Africa and *A. merus* in East Africa. Later, the *A. quandriannulatus* was subdivided further into *A. quandriannulatus* A (in South Africa) and *A. quandriannulatus* B (in Ethiopia) because the cross-mating experiments between these two species produced sterile males and displayed extensive asynapsis in the ovarian polythene chromosomes suggesting that they were different *Anopheles* species (27). Most recently, the two species were renamed *A. quandriannulatus* and *A. amharicus,* respectively (28). *A. bwambae,* one of the least distributed members of the *A. gambiae* complex, is found in geothermal salty water in Bwamba County in Uganda (29, 30). *A. comorensis* [often not listed among the above eight because little is known about it] was described in populations in the Grande Comore islands in Indian Ocean (25, 31).

1.4.2 *A. gambiae s.s.* speciation into new molecular forms

Mosquito genome is organized into 3 pairs of chromosomes namely; X or Y sex chromosomes, and autosomal chromosomes 2 and 3 (25, 32). Each of the autosomes has two ‘arms’ connected at the centromere - the longer one named right (R) and shorter one is left (L) (25, 32). Based on the analyses of *A. gambiae s.s.* polytene chromosomes (i.e. chromosomes appear thick and correspond to different densities) in the adult females’ ovarian nurse cells especially on the 2R chromosome, five configurations of paracentric (outside the centromere) inversions (32) were described [Reviewed in (33)] (**Table 1-1**). These are the j, b, c, u and d inversions, and in the wild type status, where no inversion occurs, is indicated by a positive sign (+). The j, b, c, u are non-overlapping, while the d overlaps with the u. Based on these five inversions and the wild type chromosomal forms, 12 main karyotypes were observed: +++++, jb+++,

jbcu+, jb++d, j+cu+, j+++d, +bc++, ++cu+, +bc+d, +bcu+, +b+c+, and +b++d. Accordingly, five geographical subdivisions of *A. gambiae s.s.* chromosomal forms in West Africa, namely; Bamako, Bissau, Forest, Mopti, and Savanna, were described based on patterns of inversions on the chromosome 2 (2R-j, b, c, d and u, and 2L.a) (**Table 1-1**) (*25*, *33*, *34*).

Table 1-1. Five subdivisions of chromosomal forms.

Form	Inversion karyotype	Geography
Bamako	Fixed j+cu+ and jbcu+	Bamako in Mali, north Guinea, along the Niger river.
Bissau	+++++ and ++++d	Gambiae. Restricted to West Africa
Forest	+++++, sometimes with single inversion of b, c, u or d	Associated with wetter ecological niches in Africa.
Mopti	+++++, +bc++, and +++u+ in 2R, and nearly fixed 2La	Predominate in drier habitats in Mali, Guinea, Cote d'Ivoire, and Burkina Faso. They breed throughout the year and are associated with flooded/irrigated fields.
Savanna	High frequencies of +bc++, ++cu+, +bc+d, +bcu+, +b+u+, and +b++d in 2R	Most widespread across sub-Saharan Africa.

Although, the reproductive barriers among the species in the *A. gambiae* complex exist even in sympatric species populations, extensive introgressions between the species have been documented (*8*, *33*, *35-40*). Cross talks between introgression, reproductive isolation and adaptation to ecological habitats may lead to new species and/or change in vectorial capacity i.e. how efficient the vector becomes in transmitting the malaria parasite, hence complicating the malaria transmission (*32*). For instance, the *A. gambiae s.s.* in West Africa underwent speciation into two molecular forms formerly named 'M' and 'S' for Mopti and Savanna, respectively (*32*). The 'M' and 'S' incipient species have since been renamed as *A. coluzzii* and *A. gambiae s.s.* respectively (*28*).

1.4.3 Feeding and breeding preferences of the *Anopheles* mosquito species

The *A. gambiae s.s.* are the most dominant and efficient malaria vectors in sub-Saharan Africa (**Fig. 1-3**, **Fig. 1-4**) (*3*, *25*).

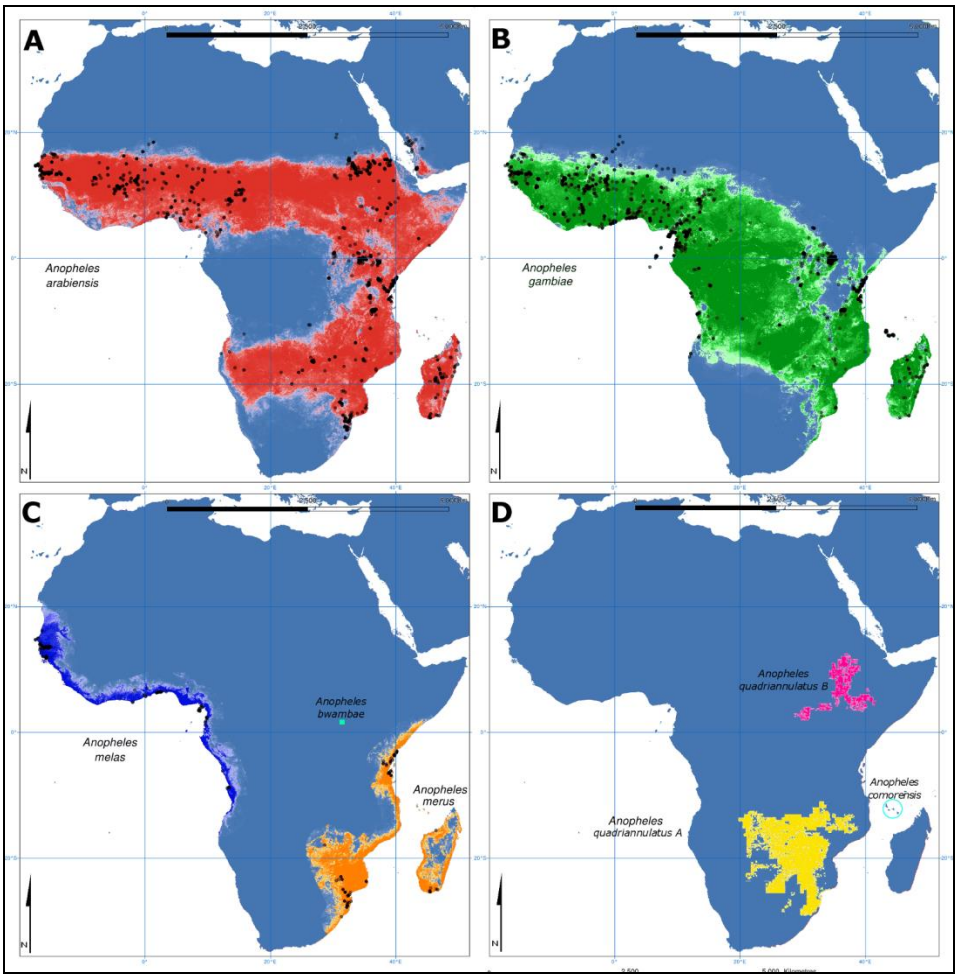


Fig. 1-3. Geographic distribution of *A. gambiae* malaria vectors in Africa.
(A) *A. arabiensis* (red).
(B) *A. gambiae s.s.* (green).
(C) *A. melas* (blue), *A. merus* (orange) and *A. bwambae* (cyan).
(D) *A. quadriannulatus* (former species A) (yellow), *A. amharicus* (former *A. quandriannulatus B*) (magenta) and *A. comorensis* (cyan circle). Figure source: Reference (25).

A. gambiae s.s. breeds mostly in the rain-dependent water pools and in fresh water puddles, whereas *A. coluzzii* shows preference for larger habitats associated with plenty of water especially from floods, rice paddies and irrigated agricultural farms (25, 41). *A. arabiensis* species can either breed in large and/or small temporal pools of water, such as those commonly found in irrigated farms (3, 25). Coastal salty or brackish water from pools, swampy and marshy form suitable niches for breeding of *A. merus* and *A. melas*, whereas other species have to adapt to such environments in order to co-exist together (25, 42-44).

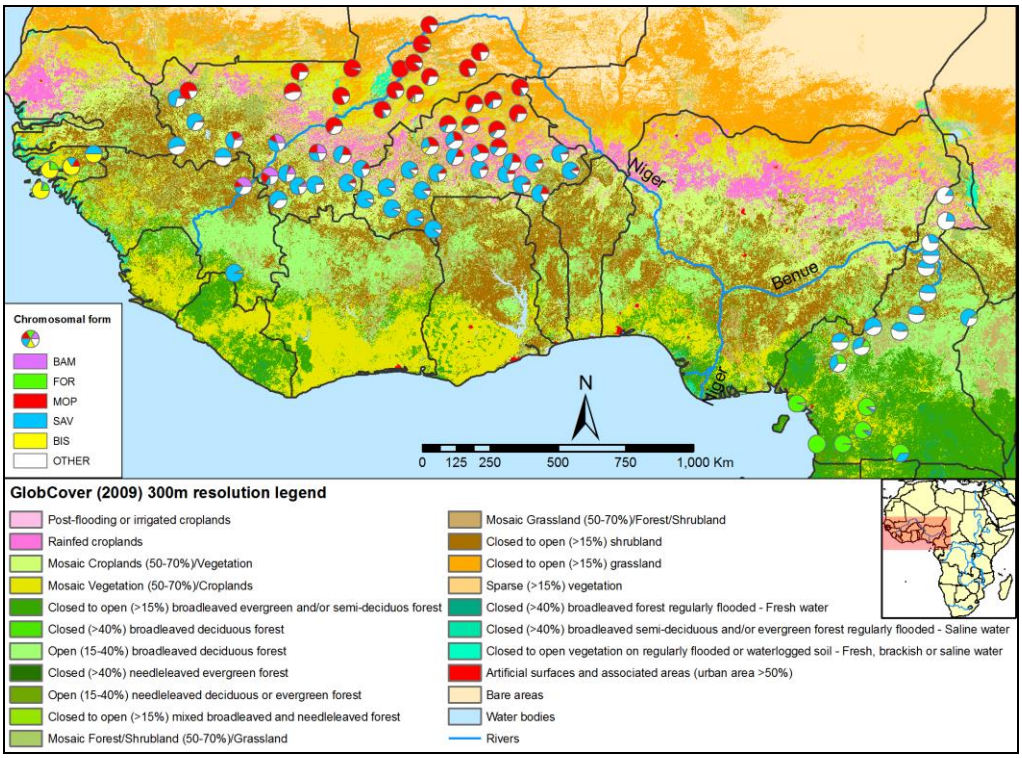


Fig. 1-4. Geodistribution of *A. gambiae* chromosomal forms in West and Central Africa. *A. coluzzii* (former M form) and *A. gambiae s.s.* (former S form) in West and Central Africa. Abbreviations: BAM; *Bamako*, FOR; *Forest,* MOP; *Mopti;* SAV; *Savanna,* and BIS; *Bissau.* OTHER represents karyotypes that could not fall into any of the above chromosomal forms. Figure source: Reference (25).

Notably, the structured distribution of different species of malaria vector populations in Africa match specific climate zones and biomes i.e. areas of land characterized by their climate and type of vegetation (**Fig. 1-5**) (25, 45).

The *A. gambiae s.s.* breeds mostly in humid savannas and during the rainy seasons but *A. arabiensis* are adaptable to dry conditions in Sahel, arid Savannas and flooded breeding sites (33, 46). In addition, sympatric populations of *A. arabiensis* and *A. gambiae s.s.* are widely distributed in Africa with fluctuations in their population numbers depending on geographical breeding zones and seasonal patterns (46). In West Africa, *A. coluzzii* breeding zones are ecologically wide in the Sahel and transition zones, floody zones as well as during the dry spell in contrast to the *A. gambiae s.s.* whose population densities are rain-dependent (25, 46).

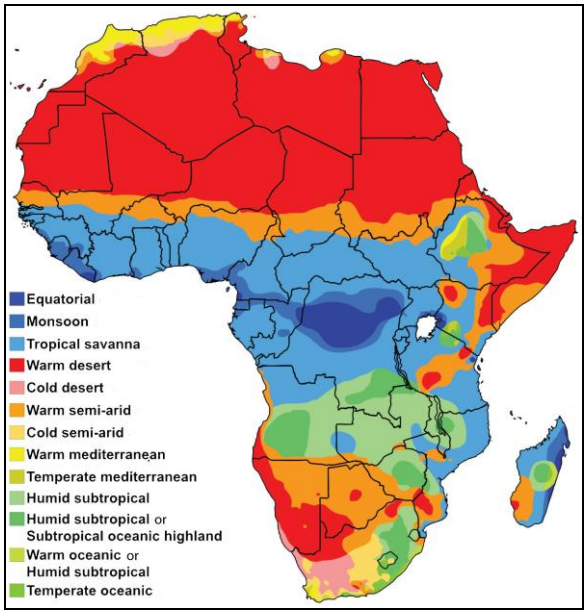


Fig. 1-5. African Climatic zones showing the ecological habitats and biomes.
In reference to **Fig. 1-3** and **Fig. 1-4**, *A. arabiensis* and *A.gambiae s.s.* inhabit considerably different climatic biomes and zones, including equatorial and humid tropical. *A. arabiensis* does well in semiarid environments. *A. coluzzii* species are confined to far-West, West and Central Africa climate zones. *A. coluzzii* also inhabits semi-arid climate zones in West Africa. Salty water breeders; *A. melas* and *A. melas* are restricted to coastal biomes in West and East Africa respectively. In general, the climate zones of equatorial/monsoon, tropical dry savanna and warm semi-arid provide key mosquito ecological niches.
Figure source: Koeppen Climate classification.

Generally, unlike *A. merus* and *A. coluzzii* species, both *A. gambiae s.s.* and *A. arabiensis* mosquito species are considered climate-generalists as they are able to colonize most of the ecological habitats across sub-Saharan Africa (3). Both *A. gambiae s.s.* and *A. arabiensis* species are well adapted to evading predators, and they prefer to breed in open and well-lit breeding sites that provide optimal environment for larval competition and development over the other species (41, 47, 48). Additionally, *A. gambiae s.s.* is highly anthropophilic because it exclusively feeds on human blood, while *A. arabiensis* species is zoophilic (attracted to and feeds on animals), exophagic (feeds outdoors) and exophilic (independent of humans) species (3, 49, 50).

1.5 Mammalian complement system in pathogen infections

The complement was discovered by Jules Bornet in nineteenth century as a heat-labile component present in blood that augments or ‘complements’ the role of antibodies in opsonization and killing of the bacteria, hence the name complement (51). It is now understood that complement is part of the innate immune system, which labels the pathogens and mediates their destruction (51).

1.5.1 Complement proteins

The complement system in vertebrates consisting of at least 30 serum proteins that circulate in the blood (51-54). Activated proteins work together to recognize or mediate the destruction of pathogens through either lysis or opsonization of pathogens or production of inflammation mediators (54). The complement proteins were classified by

assigning letters e.g. 'C' for complement, and numbers according to the order in their discovery (*51*). For example, the first complement protein (C1), is a complex consisting of C1q, C1r and C1s zymogens, the second complement protein (C2), the third complement protein (C3) and so on. The proteins are synthesized by a variety of tissues and cells. For instance C1 is produced by the intestinal epithelium, the macrophages synthesize C2 and C4, the liver - C3, C6 and C9 and the spleen - C5 and C8. During the activation process, C2, C3, C4 and C5 are activated by a proteolytic cleavage into smaller 'a' and bigger 'b' moieties, e.g. C2 protein is cleaved into C2a and C2b fragments. The smaller protein diffuses away while the bigger protein remains attached to the surface of the pathogen, with exception of the C2 whose C2a binds to the pathogen while the C2b diffuses away.

1.5.2 Activation pathways of the complement system

Activation of the complement system (**Fig. 1-6**) occurs through one of the following pathways: i) classical pathway (CP) activated by antibodies e.g. IgM and some subclasses of IgG, bound to the antigen; ii) alternative pathway (AP) activated by microbial surface proteins; and iii) lectin pathway (LP) that is activated by lectin protein bound to the specific polysaccharide sugars (e.g. mannose) on microbial surface (*52*, *53*, *55*).

Both the AP and the LP are antibody independent and, therefore, play especially important roles in cases where the body encounters pathogens for the first time (*52*, *53*). It is important to note that all the three activation pathways share a common step that involves generation of the C3b component that plays a crucial role in the complement cascade (**Fig. 1-6**) (*52*, *53*). The C3b contains a thioester site exposed by a cleavage of a thioester-bond (*54*). In principle, the C3b uses this thioester to bind to the surface of a pathogen and acts as a C5 convertase that consequently activates the C5 into the C5a and C5b fragments. The C5b together with other complement proteins (C6, C7, C8 and C9) successively form a circular C5bC6-9 complex, the so called a membrane attack complex (MAC) on the surface of the pathogen. The MAC mediates lysis (destruction of the pathogen by piercing its plasma membrane) using perforin-like C9 domains (*52-54*).

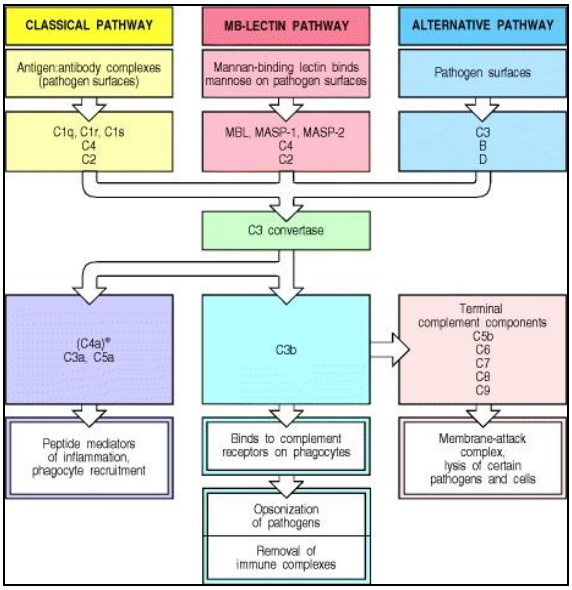


Fig. 1-6. Complement activation pathways. The activation pathways of the complement system consists of three pathways- classical pathway, lectin pathway and alternative pathway. All the pathways differ at the initiation steps, but converge at the central step-the generation of active complement protein, C3b, which forms a C5 convertase. Cleavage of C5 by a complex formed by C3b, results into C5a and C5b fragments. Both C3a and C5a meditate inflammation reactions. C5b recruits terminal components of the cascade that ultimately forms a membrane attack complex on the surface of the pathogen that damages pathogen's cell membrane. Figure source: Reference (*51*).

1.5.2.1 Activation of alternative pathway

The activation of the alternative pathway (AP) is initiated when C3b binds (from spontaneous hydrolysis of C3 in blood) to the surface of the pathogen (**Fig. 1-6**). Another plasma protein, B, binds to the C3b to form 3CbB. A plasma protease D then binds to the 3CbB and splits B to generate C3bBb and Ba. The 3CbBb molecule is an active form that remains bound to the pathogen surface, and yet at the same time acts as a convertase to cut more 3C molecules and convert them into 3Cb active forms. Another protein, called properdin or factor P, is a positive regulator that binds to and stabilizes the 3CbBb convertase hence enhancing amplification of the activation. The 3cbBb molecule can bind C3b to form a convertase that splits complement protein, C5, into C5a and C5b fragments. The smaller fragments, C3a and C5a, that diffuse away, constitute anaphylatoxins, which mediate inflammation reactions (*52*, *53*). The C5b fragment forms a complex with other complement proteins (C6, C7, C8, and C9) to form the MAC. The MAC structure anchors on the surface of the pathogen and makes a hole to lyse the bacteria. A strict control of the cascade activation to prevent MAC attacks on its own cells is ensured by a set of specialized proteins in the blood (*51-53*). For instance, further amplification of the complement cascade is stopped by membrane co-factor protein (MCP) that cleaves active form of C3b into inactive molecules on the cell surface. Factor H is another complement-regulatory protein that binds C3b to stop conversion of C5. The C3b molecule can also be cleaved to inactive form by the plasma protease factor I, with the help of cofactors such as membrane cofactors of proteolysis

(MCP or CD46) and complement receptor 1 (CR-1). Decay accelerating factor (DAF) protein found in human cells inhibits the MAC formation by destroying the assembly of C3bBb complex. Another protein called protectin (CD59) binds to the complex formed by the C5b, C6, C7 and C8, thereby inhibiting the recruitment of C9 molecules and the MAC formation. Protectin (CD59) and vitronectin (S protein) inhibit the MAC formation by binding to the C5bC6C7C8 complex.

1.5.2.2 Activation of classical pathway

The activation of the classical pathway which was first to be discovered, (*51*) is triggered when the first complement protein, C1q, binds to the pathogen surface (**Fig. 1-6**). The bound C1q together with the C1r and C1s, act as a convertase to clip C4 and C2 complement proteins. The C2 is converted into C2a and C2b, while the C4 into C4a and C4b. The C2b and the C4a diffuse away leaving a C2aC4b complex. This C2aC4b complex binds to the surface of the pathogen and acts as a C3 convertase to split the C3 protein in blood into C3a and C3b (*52*, *53*). C3b binds to the C2aC4b complex to form a bigger C2aC3bC4b complex, which acts as a C5 convertase that cleaves C5 into C5a andC5b. The terminal process is similar to the alternative pathway activation. The activation of the classical pathway is regulated by the C1 inhibitor, which binds C1r and C1s. As in the alternative pathway, the C3b molecule can be deactivated by cleavage by factor I. In addition, decay accelerating factor (DAF) inhibits the MAC formation by blocking the assembly of C3bBb complex. The C4b-binding protein (C4bBP) is a co-factor of factor I, and may block the action of C4b.

1.5.2.3 Activation of lectin pathway

The lectin pathway (LP) uses proteins similar to C1q to activate the complement reactions in the absence of antibodies (**Fig. 1-6**). These proteins include the mannose-binding lectin (MBL), which is produced in the liver and present in tissues and blood. This protein binds to a carbohydrate molecule, mannose, present only on the surfaces of many pathogens and not own cells (**Fig. 1-6**). The MBL activates the LP by binding to and activating serine proteases MASP-1 and MASP-2. MASP-2 bound to the MBL acts as a convertase that cleaves C4 and C2. The rest of the LP activation cascade is similar to those of the classical pathway, ultimately leading to MAC formation. The C1 inhibitor by binding to the MASP proteases may regulate the lectin pathway.

1.6 Insect innate immune responses against pathogens

Insects mount robust innate immune responses against invading pathogens such as viruses, bacteria (*18*), fungi (*56*) and malaria parasites in mosquitoes (*57*). The immune system is categorized into humoral and cellular defense mechanisms.

In humoral defenses, receptor molecules mediate recognition of pathogens through the activation of specific serine proteases which trigger processes such as melanization (i.e. deposition of melanin on the surface of the dead pathogen) leading to lysis or killing of pathogens. Examples of these effector molecules are *A. gambiae* antimicrobial peptides that are produced against various pathogens. These include defensin (active against Gram-positive bacteria), cecropin-1 (against Gram-positive and Gram-negative bacteria, and fungi) and gambicin (against Gram-positive and Gram-negative bacteria) (*56*, *58*, *59*).

The cellular immune responses are mediated by the mosquito blood cells (hemocytes), and include phagocytosis of the pathogens. Extensive studies in the fruit fly, *D. melanogaster* show that humoral and cellular immune responses involve three signaling pathways: the Toll pathway (*60*, *61*), Immune deficiency (IMD) pathway (*62*), and Janus Kinase/Signal Transducer and Activator of Transcription (JAK/STAT) (*63*, *64*). In *A. gambiae,* activation of Toll and IMD pathways induce transcription of effector genes (e.g. antimicrobial peptides) through NF-kappaB REL transcription factors (*62*, *65*). REL1 (analogous to *D. melanogaster* Dif) and REL2 (orthologous to *D. melanogaster* Relish) are *A. gambiae* transcription factors in Toll and IMD pathways respectively (*66*, *67*). The Toll signaling pathway in *A. gambiae* is more effective in eliminating *P. berghei* (murine) than *P. falciparum* malaria parasites (*68*). The IMD pathway also plays a role in regulation of melanization reactions, and is involved in elimination of *P. falciparum* parasites (*67*). The *A. gambiae* JAK/STAT pathway contributes to anti-plasmodial immune responses against development of early *Plasmodium* oocysts, through activation of transcription of nitric oxide synthase (NO) (*64*).

The interaction between the malaria parasite and the mosquito vector is characterized by immune defense reactions mounted by the mosquito against the development of the parasite within its body (*69*). The immune responses against the malaria parasites are broadly divided into two phases: early (first) phase- targeting ookinete stages about 18-24 h post-infection, and the late (second) phase- targeting

oocyst and sporozoite development. Malaria parasites experience the highest dramatic losses at midgut stages- the early phase of the infections, particularly the ookinetes (*4*, *5*, *70*, *71*). The immune responses against the ookinetes are mediated by *A. gambiae* complement-like proteins that circulate in the hemolymph (*4*, *5*, *18*, *71*, *72*).

1.7 *A. gambiae* complement-like system

The *A. gambiae* complement-like innate immune system is activated to attack and eliminate invading pathogens (*18*, *57*, *69*, *73*). The system was originally discovered in sea urchins (*74*) and later on, it was confirmed to be present in many other invertebrates such as ascidian, *Halocynthia roretzi* [reviewed in (*54*)] and mosquitoes (*18*, *69*). The complement-like protein components share similar sequence and structural homology (including the thioester motif) to the complement C3/C4/C5 protein components in vertebrates (*54*, *74*, *75*). In *A. gambiae,* a key protein that is activated to mediate the complement-like innate immune responses is the thioester-containing protein 1 (TEP1) in the family of TEPs (*18*).

1.7.1 Thioester containing protein 1 (TEP1)

The TEP1 was first discovered through gene known-down experiments as an opsonin that mediates phagocytosis of bacteria (*18*). Later, it was shown that TEP1 is recruited on the surface of the ookinete to promote lysis and melanization of the ookinetes, resulting in dramatic losses in parasite numbers (*4*). As research on TEP1 advanced, the protein crystal structure of the TEP1 was elucidated and found to be similar to the human complement factor C3 (*18*, *76*, *77*) (**Fig. 1-7**). Its main domains are 8 macroglobulin (MG), β-sheet CUB and α-helical thioester (TED) (*77*). The TED region protects the intramolecular β-cysteinyl-γ-glutamyl thioester bond between the TED and MG8 interphase (*18*) from spontaneous hydrolysis (*76*).

1.7.2 Activation of the TEP1 and binding to the pathogens or cells

The TEP1 is constitutively produced and secreted as a 165 kDa full-length molecule by primarily produced in the mosquito fat body, which is an equivalent of the liver in vertebrates (78). Activation of TEP1 is triggered upon septic injury or infection by pathogenic bacteria or parasites but the exact mechanism of its activation is unclear (4, 18, 78).

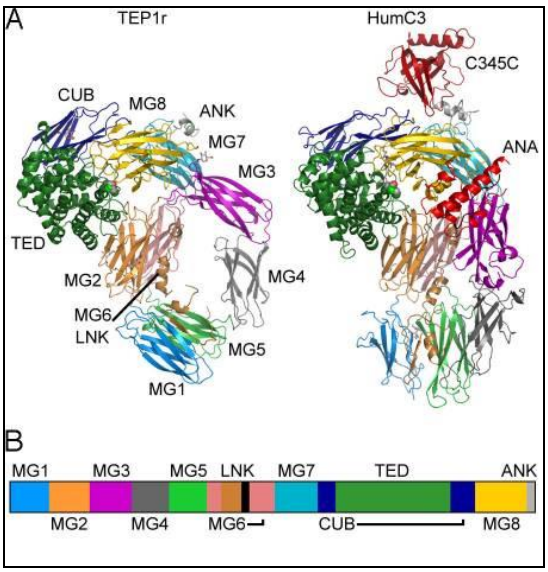


Fig. 1-7. The structure of TEP1R1. (A) Domain arrangements of TEP1R versus that of human C3. The different colors represent different domains. This show similar structure of TEP1 to the human C3. The TEP1 lacks ANA domain and hence it is less stable than the C3. (B) Schematic representation depicted different domains of the proteins as coloured in A. It consists of 12 domains in total. The interphase of MG8-TED harbours a TE, thioester bond crucial for TEP1 activation. The MG8-TED interphase also protects the TE from premature hydrolysis. Figure source: Reference (77).

During the activation process, cleavage of the full-length TEP1 generates a ~80 kDa C-terminal fragment (TEP1-C) [reviewed in (79)]. The functionality of TEP1-C is comparable to the vertebrate C3b fragment, as it carries a similar thioester site (71).

Like the mammalian C3b, the TEP1-C carries an exposed thioester site that covalently binds to: i) the surface of the Gram-negative and Gram-positive bacteria and mediates phagocytosis (18); ii) the surface of *Plasmodium* ookinetes, where ookinete-bound TEP1-C mediates a powerful anti-parasitic immune response resulting in killing and clearance of dead parasites through lysis (for *Plasmodium* susceptible mosquito strains) or melanization (for *Plasmodium* refractory mosquito strains) or both (4); and iii) the surface of damaged sperm cells and clears defective sperm cells during spermatogenesis, thus, promoting high fertility in males (80).

1.7.3 TEP1 immune responses against the invading ookinetes

TEP1 binds to the surface of *P. berghei* ookinetes between 24 and 48 h post infection (71). Heme peroxidase 2 (HPX2) and NADPH oxidase 5 (NOX5) are key enzymes that modify the surfaces of the invading ookinetes in order to 'mark' them as

targets for the TEP1 binding (81). TEP1 binding on the surface of the parasite is mediated by the interaction of TEP1 with two other proteins from leucine-rich repeat (LRR) gene family: leucine-rich repeat immune molecule 1 (LRIM1) and the *Anopheles Plasmodium*–responsive leucine–rich repeat 1C (APL1C) (82, 83). LRIM1 and APL1C protect TEP1 from premature activation (83). When either the LRIM1- or the APL1C-depleted mosquitoes were infected with *P. berghei* parasites, two observations were made: (i) in the hemolymph, the TEP1-C fragment got depleted, and (ii) binding of the TEP1-C to the parasite was abrogated (83). The family of *APL1* genes consists of *APL1A, APL1B* and *APL1C* members located in chromosome 2L. Of these, only the *APL1C* is responsible for the elevated *P. berghei* oocyst loads in the *APL1C*-depleted mosquitoes (84).

Major parasite losses within a mosquito take place during the ookinete stage, making this stage one of the most promising targets for controlling malaria transmission (21). There is no evidence that TEP1 acts on the subsequent developmental stages- the oocysts and the sporozoites (71).

1.7.4 *TEP1* polymorphism

TEP1 gene is on 3L chromosome and it is exceptionally polymorphic, coding for the 1338 amino acid long protein (5). *TEP1* allele variants are broadly grouped into two classes with two subclasses: refractory *TEP1*R* (**R1* and **R2*) and susceptible *TEP1*S* (**S1* and **S2)* (5). These *TEP1* allelic subclasses are found both in laboratory strains and in natural mosquito populations in sub-Saharan Africa (5, 8, 35). The allelic subclasses are distinguished by fixed allele-specific nucleotide and amino acid sequence variation present mainly at the thioester domain (*TED)* region (5, 8, 76, 77). Importantly, two hypervariable loops; the pre- α4 and the catalytic loop in the *TED*, have amino acid substitutions that clearly segregate the **R* and the **S* alleles. *TEP1*R* is further separated into *TEP1*R1* and *TEP1*R2* alleles mainly by five amino acid substitutions; T919G, V936A, N937K, V946M, and C1142S (35). These amino acid substitutions are located in the hypervariable loops of *TEP1* alleles and potentially affect TEP1 reactivity, binding and, thereby, functional variation of TEP1 towards pathogens, including *Plasmodium* parasites (76). Indeed, polymorphism i.e. sequence variation within a population, in *TEP1* alleles was correlated with the phenotype variation in *Plasmodium* resistance and in male fertility (5, 35).

1.7.5 *TEP1* genotypic and phenotypic variation in *Plasmodium* infections

The susceptibility and the resistance to *P. berghei* infections correlate with the *TEP1*S* and **R* mosquitoes, respectively (5, 71). Transcription of the *TEP1* gene is up-regulated within 24 h post-*Plasmodium* infection leading to activation of TEP1 for binding the invading ookinetes (71). The binding kinetics of TEP1 to the ookinetes in the *TEP1*R* mosquitoes are faster, and higher number of ookinetes are melanized than in the *TEP1*S* mosquitoes (71). Silencing the *TEP1* gene by RNA interference, promoted higher number of surviving *P. berghei* oocysts (4, 5) [reviewed in (71)].

To directly correlate phenotypes of *TEP1* genotypes bearing **R1* or **R2* or **S2* alleles, Blandin *et al.* (5) conducted *P. berghei* infections in laboratory-bred mosquito intercrosses between two mosquito lines. They showed that **R1/R1* mosquitoes were the most resistant (>80%) and melanizing (>60%), while the **S2/S2* genotypes were the least resistant and melanizing (>18%), and the **R2/R2* genotypes and all heterozygote mosquitoes portrayed intermediate phenotypes. White *et al.* 2011 (35) also observed melanization (100%) and the resistant phenotype of the *TEP1*R1/R1* mosquitoes to *P. berghei* infections. However, the infection experiments of the *TEP1*R1/R1* homozygous with human *P. falciparum* parasites were not successful due to strong selection by the mosquitoes against the *TEP1*R1/R1* genotypes. Instead, they observed lower numbers of *P. falciparum* parasite oocysts in *TEP1*R1/S* heterozygote than in the *TEP1*S/S* mosquitoes suggesting indirectly that the *TEP1*R1* allele was more resistant than the *TEP1*S* alleles, but the phenotypes were not compared at the level of *TEP1*S1* and **S2* alleles (35).

1.7.6 *TEP1* genotypic and phenotypic variation in male fertility

Recently, Pompon and Levashina (2015) reported a new role of the anti-*Plasmodium* TEP1 complement system during spermagenesis (80). They demonstrated that TEP1 is present in the testis of *A. gambiae* and mediates the efficient removal of damaged sperm cells leading to higher male fertility. In addition, silencing the LRIMI and HPX2 proteins results in disappearance of TEP1-positive spermatogonia without affecting the TEP1 expression in the hemolymph. This observation demonstrated that the LRIMI and HPX2 proteins are required for TEP1 binding to the spermatogonia, and that the complement-like cascade regulates the binding of TEP1 to the damaged sperm cells. Interestingly, comparison between homozygous *TEP1*R1, TEP1*S1* and *TEP1*S2* mosquitoes revealed that the highest degree of the male fertility correlated

with the homozygous *TEP1*S2* mosquitoes, suggesting that male fertility is dependent on *TEP1* polymorphism i.e. allelic variation. The role of the susceptible *TEP1*S2* allele in reproduction may underlie one of the reasons why mosquitoes maintain the susceptible alleles in their populations, and this, could have important consequences for malaria transmission. Therefore, the evolution of the *TEP1* locus and genetic factors shaping resistance to malaria parasites may be a consequence of *TEP1* pleiotropic trade-off between *TEP1* allelic fitness in immunity and reproduction.

1.7.7 Ecological significance of evolutionary forces

1.7.7.1 Concept of population genetics

Maintenance of genetic variation in gene loci in mosquito populations, for example in the *TEP1* locus (5-8), provides insights into biological processes and evolutionary forces underlying functional traits in natural populations (85). The change in the DNA sequences caused either by mutations or by genetic recombination brings about the genetic variation (6, 86, 87). Mutations are sources of new alleles or genes, and are more frequent and more beneficial in unstable environments (88). Genetic recombination on the other hand occurs through new allele combination (88). It is the source of most of the genetic variation in a population and the main source of variation underlying gene evolution. The phenotypes or traits that are manifested in an individual organism are defined by their genotypes, or their genetic makeups, and may depend on the interactions of different genes and the environments (85). These aspects form the basics of the population genetics i.e. the study of genetic variation within populations, through the manipulation of allele and genotype frequencies from one generation to another (85). It also deals with the study of various forces that bring about evolutionary changes in species populations over time (85). In this context, it provides an understanding of ecological and evolutionary processes and impacts of demography on local populations, gene frequencies and phenotypes (85, 89).

1.7.7.2 The Hardy-Weinberg principle

One of the basic models in population genetics is the Hardy-Weinberg principle or the Hardy-Weinberg equilibrium (HWE), [named after the English mathematician G.H. Hardy (1877-1947) and the German physiologist W. Weinberg (1862-1937) who, in 1908 independently and concurrently formulated the principle] (85) (90). The HWE is used to deduce theoretical genotype frequencies and infer certain evolutionary processes of a given population (85, 90). The principle assumes that in sexually reproducing and

non-evolving population evolutionary forces are absent or are negligibly small in magnitude, the alleles or genotypes remain constant from one generation to the next (88, 90). Let us, for example, hypothetically consider *TEP1* locus with two alleles, R and S, in a given mosquito population. The allele diploid combination under random mating gives RR, RS and SS genotypes. In the first generation, the allele frequencies of the R and S alleles may be denoted as p and q, respectively. Summation $p + q = 1$ (100%). The mathematical relationship for the expected genotype frequencies is given by $(p\text{R}+q\text{S})^2 = p^2$, $2pq$ and q^2 for RR, RS and SS, respectively. Hence $p^2\text{RR} + 2pq\text{RS} + q^2\text{SS} = 1$ (100%). In the second generation, the frequency of say R remains the same that is p. Therefore, $p^2 + pq \,/\, p^2 + 2pq + q^2 = p\,(p+q) \,/\, (p+q)^2 = p \,/\, p+q$, hence $p = p$ as in generation 1.

The principle of the HWE is based on the following assumptions (88, 90):

1) No mutations - no random change in base sequence of the genetic material within individuals. These changes are heritable and result in genetic variation, but they occur rarely with the majority being harmless or neutral (91).

2) No migration - no movement of individuals into and out of the population. However, this does occur when some new comers enter into a population or some individuals move out of the population, hence causing successful movement of alleles i.e. gene flow, and genetic variation (91).

3) Large (infinite) population. This provides all the possible kinds of zygotes to be formed in frequencies determined by the gametic frequencies (92). But bottlenecks and founder effects often occur (92).

4) Random mating - no sexual selection. Mating should not be determined by any preferences associated with specific genotypes (92). But most animals mate selectively and may have differential mating success among individuals (91).

5) No selection. Thus, all genotypes should have equal reproductive ability (92). But genotypes are not equally adaptive, hence natural selection does happen (86).

Additional assumptions include (90);

6) The organism is diploid, and is equally fertile to produce gametes according to the frequency of parents (92).

7) Generations are non-overlapping. Gene locus under consideration has two alleles. But other genes such as *TEP1*, are multiallelic (5).

9) Allelic frequencies in both males and females are identical. This means that the gene frequencies in both males and females are the same (92).

10) Reproduction is sexual, with equally fertile gametes to have equal chances of becoming a zygote (92).

When any of the selection forces operates in the population, the frequencies of alleles and genotypes in the population change from one generation to the next. Significant deviation from the HWE expectation shows that selection is happening in the population (88). Hence the population is undergoing evolution resulting in selection for fitter individuals (**Fig. 1-8**) (86, 93).

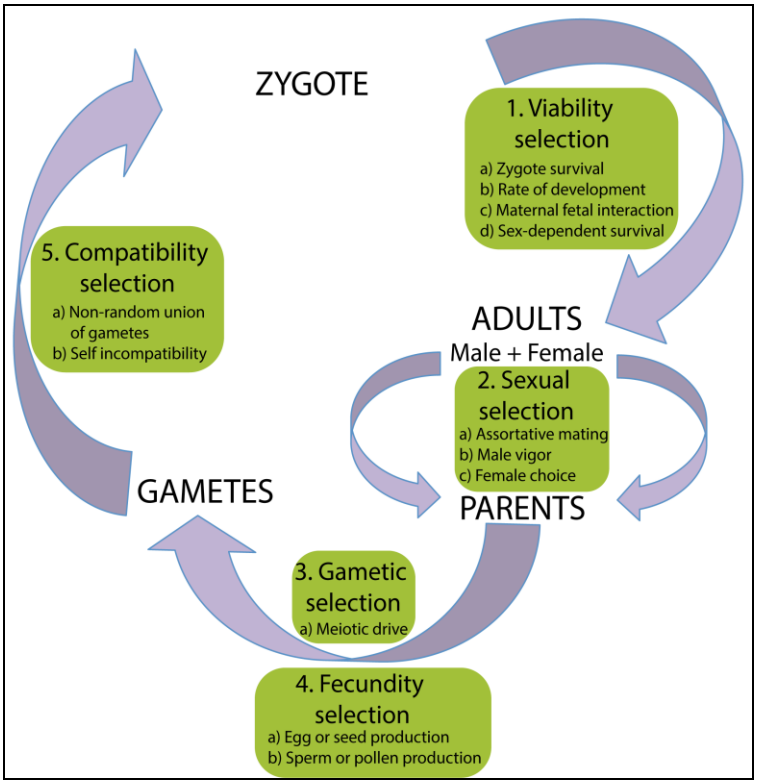


Fig. 1-8. Selection forces acting on the life stages of an organism. Selection in a sexually reproducing diploid organism occurs in many stages in the life cycle. Fitter individuals are selected to survive through to the next developmental stages of population (86). Between each developmental stage, there are selection bottlenecks that determine the survival of fitter individuals to go the next life stage. These selections are classified into viability, sexual, gametic, fecundity and compatibility selection forces (86, 93). Figure source: Original to this thesis.

Although its applicability is not universal, the HWE principle provides an important platform to form a hypothesis about genetic structure of a population and design experiments in population genetics including mosquito populations (86, 88, 90). Given the genotype frequencies and the number of individuals per genotype in a population, it is possible to use these parameters to test for the HWE deviation by calculating the expected genotype frequencies (90). A commonly applied test in these analyses is the standard Chi-square (χ^2) test, which is calculated as follows;

$$\chi 2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

In cases where the expected genotypes are high for some genotypes and very low e.g. 5 individual, the following conservative χ^2 test can be used;

$$\chi 2 = \sum \frac{(|Observed - Expected| - 0.5)^2}{Expected}$$

Where, absolute (i.e. negative signs are ignored) difference between the observed and the expected genotypes is calculated, and 0.5, which is meant to reduce the χ^2 value is subtracted first before taking the square (90).

The χ^2 and number of degree of freedom (*df*) associated with χ^2 value provide a number for assessing goodness of it. The *df* is given by;

df = Number of classes of data – Number of parameters estimated from the data.

The *df* for a multi-allelic locus in a population is modified to capture complexity of alleles and genotypes analyses, where in general, *df* is given by the difference between the number of possible genotypes (*m*) and the number of alleles (*n*) present at the locus in the population, i.e. $df = m\text{-}n$ (90).

The expected and the observed genotype frequencies provide parameters to test inbreeding and outbreeding in populations. Inbreeding within population occurs when organisms which are related or identical by descent (e.g. same genotype) mate together (94). The inbreeding coefficient (*F*) – the probability that two alleles at a locus in an inbred individual are identical by descent. It takes into account immediate inbreeding and reveals whether the population is inbred or outbred (decrease in homozygosity) (88, 94). Decrease in homozygosity occurs when mating between different genotypes leads to increase in heterozygosity commonly referred to as 'isolate is breaking' or the Wahlund principle (94). Between two or more populations, the comparison forms the basis of population fixation index (F_{ST}) (94). The F_{ST} considers only the autozygosity i.e. genes in homozygote as a mating between related individuals, occasioned by population subdivision and gives a measure of population subdivision that indicates the proportion of heterozygosity found between populations relative to the amount within populations (89, 94). Another *F* statistics, F_{IT}, unites both inbreeding and population structure, and gives the probability of autozygosity of an inbred individual relative to the whole population, where all the subpopulations combine and undergo random mating (95).

1.7.7.3 Natural selection at the *TEP1* locus and other *TEP* loci

Selective pressures that drive evolution may come from the environment itself, either naturally (non-random natural selection) or by chance (random genetic drift) (*6*, *96*, *97*). Ecological niches are sources of selection forces that act on standing genetic variation at gene loci in natural populations (*98*, *99*). This enables adaptation of *Anopheles* populations to different breeding habitats. Immune genes such as *TEP1*, are functionally constrained under evolution by predominantly purifying selection (*6*), and so detecting evidence of sites under positive selection is difficult (*100*).

White *et al.* (*35*) identified candidate gene targets underpinning divergent speciation and selection in *A. coluzzii* and *A. gambiae s.s.* natural populations. To evaluate the extent of the divergence between these two species, they complemented genome-wide scanning in paired field-caught local mosquito populations from Mali with re-sequencing and genotyping samples from West, Central and East Africa. Marked divergence on chromosome 3L was observed between the two species, underscoring putative gene [AGAP010817] and known [*TEP1, TEP3*, *TEP10*, *TEP5* (annotated as *TEP11* in Vector Base)] immune genes. In *TEP1, TEP3* and *AGAP010817*, the divergence was uniform within all the *A. gambiae s.s.* across all the countries of study suggesting that *A. gambiae s.s.* geographic scope for these genes was limited at this genomic regions. Population pairwise-comparison between geographic sites was significantly high in Mali and Burkina Faso, but indistinguishable between *A. coluzzii* and *A. gambiae s.s.* populations in Ghana and Cameroon suggesting geographic differences in mosquito species and genomic regions.

Interestingly, the *TEP1* locus exhibits the most extreme divergence due to the existence of multiple *TEP1* alleles: *A. coluzzii *R1* (Mali and Burkina Faso), *A. coluzzii *S* (Ghana and Cameroon) and *A. gambiae *S* (Mali, Burkina Faso, Ghana) alleles. Geographic- and species-restricted forces correlated consistently with the *TEP1* allelic variants. For instance, the **R1* allele was geographically restricted to Mali and Burkina Faso at near fixation within *A. coluzii* local populations i.e. presence of the *TEP1*R1/R1* homozygotes, and the **R1* gene flow was spreading to Ghana. On the other hand, the *TEP1*S* and **R2* alleles were detected in all the *A. gambiae s.s.* populations.

In addition, microarray scans of *TEP1, LRIM 1* and *APL1C* gene loci between *A. coluzzii* and *A. gambiae s.s.* populations sampled from West and Central Africa, revealed that *TEP1* locus was under significant ecological pressures of divergence and

speciation (35). The authors hypothesized that selective forces act on the *TEP1* locus and contribute to the adaptive divergence in the *A. gambiae* species, whereas another study did not find significant evidence of selection acting on the *LRIM1* and *APLC1* gene loci (101), even though their gene products interact together as a complex to stabilize and promote TEP1 binding to the pathogen surface. Indeed, pathogens, demographic and other ecological factors that are present in the larval breeding sites, and/or the vector species, are implicated as key contributors to the *TEP1* allelic diversification, and natural selection driving forces of adaptive divergence of local malaria vector populations at the *TEP1* locus *(6, 8, 35, 102, 103)*.

1.7.7.4 Contribution of recombination to the *TEP1* gene diversity

It is important to note that within any given gene locus, different sites are not under equal or similar selection forces (6). Obbard *et al.* (6) showed that the *TEP1* locus is a target of natural selection and that the hallmark of underlying *TEP1* genetic variation is the independent gene conversion events due to recombination (genetic exchange) between *TEP1* locus and *TEP11(5)* and *TEP6* loci. For instance, the *TEP1* chimerical sequence similarities with the *TEP5* and the *TEP6* were exemplified by a *TEP5*-like portion upstream of *TEP1*, and a *TEP6*-like region towards 3´ end of the *TEP1* gene. Moreover, the *TED* of *TEP1*R* carried a *TEP6*-like portion, which overall translated to *TEP1*R* (3.2%) and *TEP1*S* (14.6%) divergence with the *TEP6* gene, suggesting a more recent ancestry of *TEP1*R* than *TEP1*S,* to the *TEP6* (6). In summary, the data suggest distinct histories of *TEP1* allelic variants. Therefore, in addition to natural selection, the high divergence and functional polymorphism between the *TEP1* alleles could have been driven by gene conversion mechanisms.

1.8 Research gaps

Understanding the biogeographic distribution of *TEP1* allele/genotypes in *A. gambiae s.l.* offers an opportunity to infer their impact on malaria epidemiology and to inform future research directions, and malaria control stategies and policies (*104*).

Targetted-control measures (*105-107*) as well as climate change (*48, 108-111*) may significantly modulate abundances of the local mosquito species populations. This may promote selection of efficent malaria vectors which could be resistant or susceptible to insecticides or pathogens including human malaria parasites. Having genetic markers that offer high throughput genotyping of local *A. gambiae s.l.* mosquito populations across Africa are of great importance, especially for monitoring dynamics in the mosquito populations that may have implications in malaria transmission.

Previous genotyping methods provided clear distinction, and species and geographic distribution ranges between the *TEP1*R1* and *TEP1*R2* allelic classes (*8, 35*). However, these studies were unable to distinguish between the *TEP1*S1* and the *TEP1*S2* alleles. Therefore, the species and biogeographic distribution ranges of *TEP1*S1* and the *TEP1*S2* alleles across Africa remains largely unclear. Moreover, phenotypic differences between *TEP1* alleles with respect to infections by human malaria parasite are unknown.

1.9 Aims of the thesis

This thesis aimed at **a)** characterizing *TEP1* alleles/genotypes of local *A. gambiae s.l.* populations sampled across four countries in sub-Saharan Africa (Mali, Burkina Faso, Cameroon and Kenya), and **b)** assessing the impact of *TEP1* variability on *P. falciparum* development. The following specific research questions were formulated in order to achieve the research aims:

1. Identification of single nucleotide polymorphisms (SNPs) within the *TEP1* locus that could be harnessed as genetic markers for high throughput genotyping;
2. Distribution and frequencies of the *TEP1* genotypes and alleles across Africa;
3. Structuring of *TEP1* genotypes along the ecological sites and geographic regions;
4. Forces that shape distribution of *TEP1* genotypes; and
5. Impact of *TEP1* variability on resistance to human *P. falciparum* and murine *P. berghei* parasites.

The findings of this thesis are expected to guide the formulation of hypotheses that may motivate future research on selection and roles of different *TEP1* alleles and genotypes in the natural vector populations. The high throughput *TEP1* genotyping strategies as used in this study, may be adopted and incorporated in the malaria vector control programs for routine genotyping of the local malaria vector populations across Africa. This will offer prediction of vector population dynamics due to the impact of human activities and climate change which may influence malaria transmissions.

1.10 Outline of the thesis

This thesis is divided into four chapters:

- **Chapter 1** gives background information on malaria and its transmission cycle. It reviews the relevant literature on the current knowledge and challenges on the dynamics of distribution of malaria vectors in sub-Saharan Africa, and advances on the innate immunity of the malaria mosquito towards the malaria parasites. It identifies research gaps and provides the aims of the thesis;
- **Chapter 2** reports development of a high throughput PCR-RFLP genotyping approach that can be used to sufficiently identify all *TEP1* alleles. It provides and discusses data on genetic diversity and biogeographic distribution ranges of *TEP1* alleles/genotypes of local malaria vector populations in Mali, Burkina Faso, Cameroon and Kenya;
- **Chapter 3** provides insights into the genetic and phenotypic variation of *TEP1* alleles or genotypes and their impact on development of human- and murine-malaria parasites. It highlights challenges encountered in breeding mosquitoes bearing the *TEP1*R1* alleles, and suggests future breeding strategies that may be explored in a bid to enhance successful routine rearing of such mosquitoes under the laboratory and/or field conditions; and
- Further, **Chapter 4** gives the general discussions and proposes hypotheses based on the results of this thesis. It suggests open questions that remain to be addressed by future studies and further wraps up with the perspective of the study, implications/conclusion of the key findings.

Chapter 2

Biotope-specific factors shape *TEP1* genetic variation in the populations of *Anopheles gambiae* across sub-Saharan Africa

Biotope-specific factors shape *TEP1* genetic variation in the populations of *Anopheles gambiae* across sub-Saharan Africa

2.1 Summary

Species of the principal malaria vector *A. gambiae s.l.* have adapted to a wide range of ecological niches (biotopes or habitats) across Africa (112). Many studies based on mtDNA, paracentric chromosomal inversions and microsatellite markers have described genetic differentiation, local adaptation and gene flow in vector populations in sub-Saharan Africa, (113-117). Indeed, putting in place effective vector control measures to curb malaria transmission requires a wider understanding of genetic variation and selective pressures underlying structuring of vector populations, and malaria transmission. Here, *TEP1* locus was used to examine genetic architecture of field vector populations. The study developed and used a PCR-RFLP genotyping method to identify *TEP1* genotypes and allelic subclasses. We identified generalist (**R2/R2*, **R2/S1* and **S1/S1*) and specialist (**R1/R1*, **R3/R3, *R3/S1, *S2/S2* and **S1/S2*) *TEP1* genotypes. We show that **R2* and **S2* are the most conserved alleles suggesting that they may represent ancestral alleles that have been maintained over generations. The contribution of intronic polymorphism to the evolution of *TEP1* alleles and genotypes is discussed. These findings suggest a trade-off between intrinsic forces maintaining ancestral genetic polymorphism and extrinsic factors that drive vector adaptation to local ecological ecotypes.

2.2 Introduction

Physical isolation and premating reproductive barriers limit reproductive interactions between malaria populations breeding in distinct habitats (118, 119). An example of geographic separation is the Great Rift Valley (GRV), a massive trench that stretches from North to South of Kenya separating it into three ecological strata: (i) West; (ii) along the GRV trench; and (i) East of the GRV. The GRV acts as a gene flow barrier in the geographical locations between the western, within the valley and the eastern populations leading to genetic and reproductive isolation (113, 120-122). Changes in ecological circumstances due to human activities and climate change affect ecological abundance of vector populations such as constriction or expansion of ecological niches (105, 123). Climate change is predicted to result in rise of sea water

levels forcing the mosquito species living in the salty habitats to move and colonize new fresh-water habitats, and this may have consequences for malaria transmission (*124*).

Sympatric mosquito species have reproductive restrictions such as premating and post-mating barriers between them (*34*, *125-127*). Ecological factors driving these reproductive restrictions vary from one habitat to another, thus relaxing the gene flow barriers between divergent species leading to incomplete reproductive restrictions (*99*). For instance in West Africa, natural hybrids between *A. coluzzii* and *A. gambiae s.s.* in sympatric populations (*8*, *35*, *99*) do occur mostly in low frequencies (<20%) suggesting porous reproductive restrictions (*8*, *39*, *115*, *128-131*). Similar observations on hybridization have been reported for *A. gambiae s.s.* and *A. arabiensis* (*36*).

Genetic structure of a population can be described based on chromosomal arrangement, microsatellites and mtDNA. There is a correlation between local adaptation to different climatic zones and paracentric chromosome inversions in mosquito populations (*32*, *119*, *132*, *133*). As such, higher and lower inversion frequencies correlate with arid/dry and wet ecotypes, respectively. Moreover, peaks and troughs match dry and rainy seasons, respectively (*32*, *132*). On an exploded geographic picture, Lehmann *et al.* (*134*) described population structure of mosquitoes across ten African countries using 11 microsatellite markers and observed genetic differentiation into two substructures; Northwest and Southeast population groups.

The *A. gambiae* thioester-containing protein 1 (TEP1) encoded by a polymorphic gene (i.e. bearing different alleles-**R1*, **R2*, **S1* and **S2*), plays a role in antimicrobial and anti-malarial immune responses, and in male fertility (*4*, *5*, *18*, *77*, *80*, *135*). Trade-off between reproduction and immunity could have a direct consequence on mosquito population abundances and malaria transmissions. This makes *TEP1* locus a promising target for new malaria control interventions (*6-8*, *78*). *TEP1* locus is under selective pressure for adaptive divergence in West and far-West Africa leading to geographic differences in *TEP1* allele frequencies (*6*, *8*, *35*). However, *TEP1* genotyping methods used did not distinguish between **S1* and **S2* alleles, thus biogeographic distribution ranges of these alleles remains unknown (*6*, *8*, *35*, *80*). Largely, due to lack of high throughput *TEP1* genotyping methods, it is unclear how *TEP1* genetic diversity influences the genetic structure and local adaptation of natural vector populations, and development of human malaria parasites (*6-8*).

In this chapter, I explored the suitability of the *TEP1* locus for genotyping the main malaria vector *A. gambiae* to dissect patterns of distribution of malaria vector species and *TEP1* genetic diversity. *A. gambiae s.l.* species were sampled across sub-Saharan Africa, namely West (Mali, Burkina Faso), Central (Cameroon) and East (Kenya) Africa. I examined the hypothesis that *TEP1* polymorphism shapes structures of local mosquito populations. Specifically, the following objectives were addressed: i) whether single nucleotide polymorphisms (SNPs) within the *TEP1* locus could be harnessed as a genetic marker for high throughput *TEP1* genotyping; ii) patterns of distribution of *TEP1* genotypes and alleles across Africa; and iii) structuring of *TEP1* genotypes and alleles according to ecological and geographic regions.

All known *TEP1* genotypes and alleles were identified using a simple high-throughput PCR-RFLP genotyping approach (*80*). I identified interesting signatures of natural selection on both intronic and exonic *TEP1* sequences. Further, I identified a new *TEP1* allele, hereafter named **R3*, which was maintained as **R3/R3* and **R3/S1* genotypes in *A. merus* population along the coastal region of Kenya. Patterns of distribution of *TEP1* genotypes and alleles suggested new biogeographic distribution. In brief, genotypes were categorized into four groups, proposed as follows: generalist (**R2/R2, *R2/S1* and **S1/S1*) due to wide distribution in relatively high frequencies in all species and most countries, specialist (**R1/R1, *R3/R3, *R3/S1, *S1/S2* and **S2/S2*) since they were found restricted to specific species and locations, rare (**R1/S1, *R1/S2* and **R2/S2*) as they were in very low frequencies in some species in West Africa, and undetected (**R1/R2, *R1/R3, *R2/R3* and **R3/S2*) genotypes which were not found in any mosquito samples. On the other hand, alleles were grouped into two biogeographic groups; generalist (**R2* and **S1*) found in almost all species across Africa, and specialist (**R1, *R3* and **S2*) alleles which were found or restricted to species and locations in Africa.

The specialist **R1* allele and **R1/R1* genotype were identified in Mali and Burkina Faso. The *A. coluzzii* **R1* allele correlated with arid conditions, suggesting a link to resistance to desiccation and ability to withstand harsh breeding conditions of the Sahel zone in Mali and Burkina Faso. In Kenya, the emergence of new **R3* private allele may be beneficial to *A. merus* in confronting certain biotic and/or abiotic ecological constraints specific to their ecological niches.

In Cameroon, both *A. coluzzii* and *A. gambiae s.s.* had inverse proportions between **R2/S1* and **S1/S2* genotype frequencies, and between **R2* and **S2* allelic frequencies, suggesting a competition between the **R2* and **S2* alleles. The most predominant alleles and genotypes in most species were **S1* and **S1/S1* respectively, except *A. coluzzii* in Mali and in some habitats in Burkina Faso which had **R1/R1* genotypes. The heterozygote **R2/S1* genotypes were the most widespread across Africa in all species except in *A. coluzzii* in Mali and in Burkina Faso, suggesting that both **R2/S1* and **S1/S1* genotypes have the highest fitness advantage. In addition, haplotype clustering of alleles showed that the**R2* and **S1* haplotypes were the most shared between species across Africa suggesting that these alleles in the vector populations are the most conserved.

Collectively, these findings suggest that local selection factors drive vector adaptation to ecological biotopes according to mosquito species. As the distribution of *TEP1* genetic diversity matches African climatic zones, I propose that *TEP1* locus contributes to the local adaption of mosquitoes to the prevailing environmental conditions. Based on the above findings, I conclude that the genetic variation at the *TEP1* locus shapes the population genetic structure of local malaria populations. For the first time, this study mapped *TEP1* alleles and genotypes in four malaria vector species from four African countries. Accordingly, new biogeographic distribution ranges of the *TEP1* alleles and genotypes have been proposed. Therefore, structure of mosquito populations at the *TEP1* locus offers an important tool in assessing gene flow radiation and genetic dispersal of malaria vectors. I propose that the SNPs that were used for *TEP1* genotyping in this study could be harnessed as genetic markers for high throughput *TEP1* genotyping as they provide a robust approach for dissecting genetic population structure of *A. gambiae s.l.*. Policy makers may incorporate these approaches into vector-control programs for surveying, monitoring and documenting the conjectures of population dynamics in local malaria vectors, as they come in handy in forecasting future demographic circumstances that bring severe epidemiological consequences.

2.3 Material and Methods

Materials, equipment and software that were used in this chapter are listed in Appendix **1**A-D.

2.3.1 Fieldwork samples and sample origin

Four sub-Saharan African countries (Burkina Faso, Cameroon, Kenya and Mali) were chosen for sample collections due to diverse geographical locations and climatic zones, as well as availability of collaborators (**Fig. 2-1**A). Larvae (Mali, Burkina Faso, Cameroon, and Kenya), and adults [mating swarms (Burkina Faso) and indoor resting mosquitoes (Kenya)] were sampled between 2009 and 2016 (**Fig. 2-1**; **Table 2-1**).

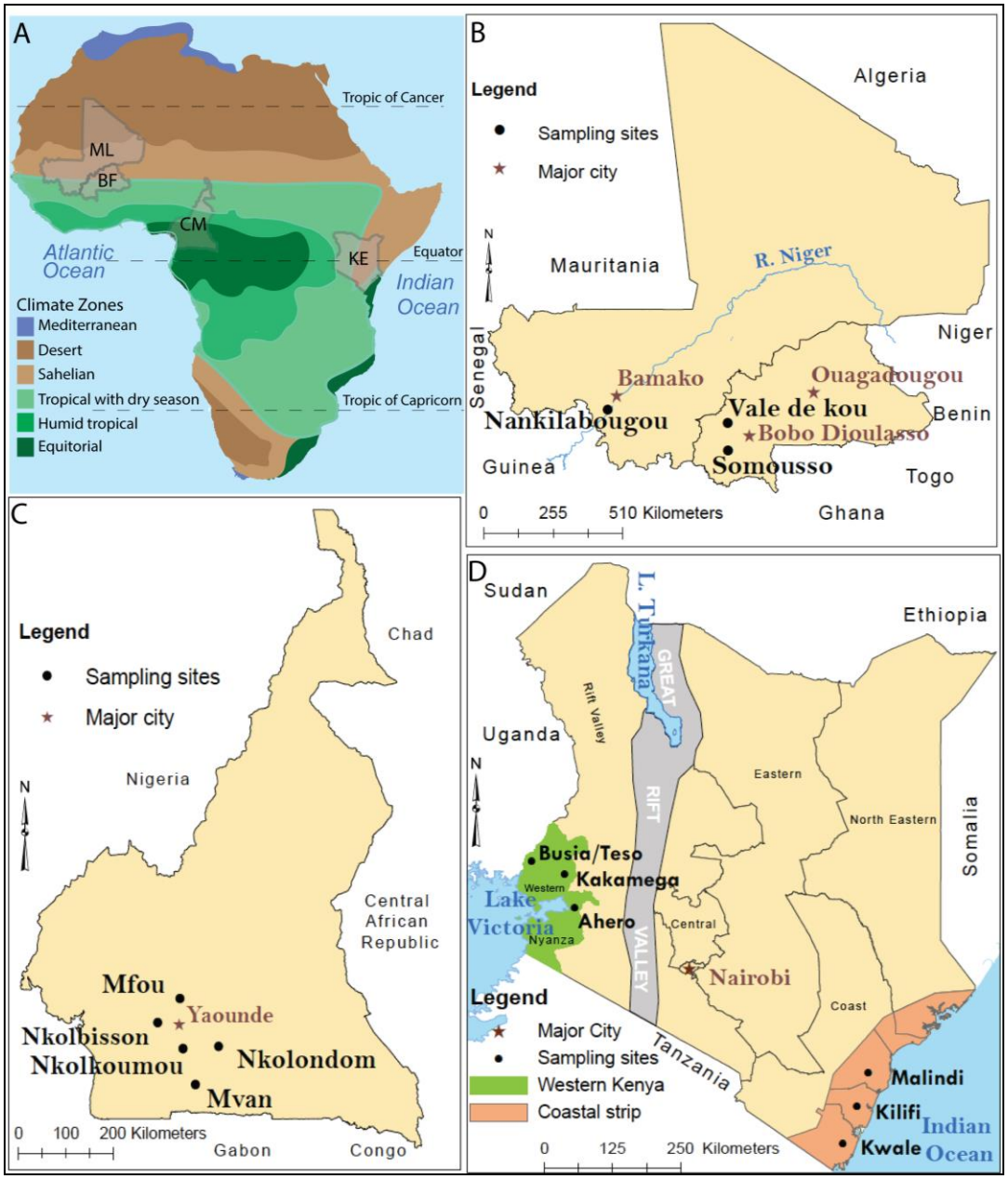


Fig. 2-1. Sampling sites investigated in this study.
(A) Overview of sampling sites in the context of the African climatic zones. Mali (ML) and Burkina Faso (BF) lie in warm savanna with a single dry season. Cameroon (CM) is in humid equatorial climate. Kenya (KE) is in warm equatorial and savanna climate zones. Locations of sampling sites in:
(B) ML and BF;
(C) CM; and
(D) KE.

Table 2-1. Information on the sampling sites.

Country	Site	Geographic coordinates	Year	Ecological domain	Landscape	Bioclimate
ML	NK	12.17 N, 8.29W	2009	Sahel	farmland	very hot semi dry
BF	VK5	11.3N, 4.4W	2009	Sahel	farmland	hot semi dry
BF	VK7	11.4N, 4.4W	2009	Sahel	farmland	hot semi dry
BF	SM	11.02N, 4.06W	2009	Savanna	savanna	very hot moist
CM	MF	3.97N, 11.94E	2009	Savanna	rainforest	very hot wet
CM	MV	3.82N, 11.53E	2009	Savanna	urban	very hot wet
CM	NS	3.88N, 11.46E	2009	Savanna	urban	very hot wet
CM	NM	3.87N, 11.4E	2009	Savanna	rainforest	hot wet
CM	ND	3.95N, 11.5E	2009	Savanna	rainforest	hot wet
KE	AH	0.17S, 34.93E	2009, 2012, 2016	Equatorial	farmland	hot moist
KE	BT	0.46N, 34.12E	2009	Equatorial	farmland	hot wet
KE	KK	0.31N, 34.79E	2009	Equatorial	farmland	hot wet
KE	KL	3.89S, 39.91E	2009, 2011	Savanna	coastal	hot wet
KE	KW	4.17S, 39.47E	2009	Savanna	coastal	hot wet
KE	MD	3.25S, 40.11E	2009, 2016	Savanna	coastal	hot semi dry

Countries: ML-Mali; BF-Burkina Faso; CM-Cameroon; KE-Kenya. **Sites:** NK-Nankilabougou; VK5-Vale de Kou 5; VD7-Vale de Kou 7; SM-Somoussou; MF-Mfou; MV-Mvan; NS-Nkolbisson; NM-Nkolkoumo; ND-Nkolodom; AH-Ahero; BT-Busia/Teso; KK-Kakamega; KL-Kilifi; KW-Kwale; MD-Malindi.

In Mali, Nankilabogou (NK) sampling site featured sympatric population of *A. coluzzii* and *A. gambiae s.s.*. This sampling site is about 66 km Southwest of Bamako in proximity to the river Niger. (**Fig. 2-1**B). Two main sites in Burkina Faso, Vale de Kou (VK5 and VK7 villages) and Somousso village, are located approximately 400 km southwest of Ouagadougou (**Fig. 2-1**B). In Vale de Kou and Somousso, *A. coluzzii* and *A. gambiae s.s.* populations coexisted in allopatry and sympatry, respectively. Unlike Somousso with temporary breeding habitats, Vale de Kou District is an agricultural rice field sourcing water for irrigation that provided permanent long-term large breeding sites for mosquitoes. In Cameroon, mosquito samples of sympatric *A. coluzzii* and *A. gambiae s.s.* mosquitoes were collected from five districts: Mfou - about 60 km Northwest of Yaounde, Mvan - 10 km South of Yaounde; Nkolondom - Southeast of Yaounde; Nkolkoumou - South of Yaounde and Nkolbisson - about 10 km Northwest of Yaounde (**Fig. 2-1**C). In Kenya, sampling was performed in two regions, western and coastal Kenya, which are separated by the GRV (**Fig. 2-1**D). In western Kenya, there were three local sampling locations: Ahero (native irrigated rice farms), Kakamega (agricultural land with tall vegetation including the forests) and Busia/Teso (a grassland in the savanna zone). Along the coastal Kenya, three locations were chosen: Kwale on South coast, Malindi and Kilifi on the North coast.

2.3.2 Species identification

The genomic DNA (gDNA) was extracted from larvae or adult legs using DNeasy kit (Qiagen, USA). The gDNA (10 ng/µl) was used in PCR standard methods to identify mosquito samples into sibling *Anopheles* species according to the published protocols (*136*, *137*). Briefly, *A. gambiae* molecular forms in Mali, Burkina Faso and Cameroon were identified by a short interspersed-PCR (*SINE-PCR*) approach, which amplifies a 200 bp insertion polymorphism at locus *S200 X6.1* located in chromosomes X (*137*). The following primers were used; forward primer 5′-*TCGCCTTAGACCTTGCGTTA*-3′ and reverse primer 5′-*CGCTTCAAGAATTCGAGATAC*-3′ which amplifies a 479 bp fragment in *A. coluzzii* (former M form) and 249 bp in *A. gambiae s.s.* (former S form) (*137*). In Kenya, *A. arabiensis, A. gambiae s.s.* and *A. merus* sibling species were identified using a multiplex ribosomal DNA-PCR method, which distinguishes the species based on species-specific sequence variation in the ribosomal DNA intergenic spacers (*136*). This method utilizes one universal primer-UN 5′-*GTGTGCCCCTTCCTCGATGT*-3′ and four species–specific primers*; A. arabiensis-AR*

5′-AAGTGTCCTTCTCCATCCTA-3′, *A. gambiae s.s.-AG* *5′-AAGTGTCCTTCTCCATCCTA-3′*, *A. melas/A. merus-ME* *5′-TGACCAACCCACTCCCTTGA-3′*, and *A. quadrianulatus-QD* *5`-CAGACCAAGATGGTTAGTAT-3′*. The primers amplify 315 bp for *A. arabiensis*, 390 bp for *A. gambiae s.s.*, 464/466 bp for *A. melas/A. merus,* and 153 bp for *A. quadrianulatus* (136).

3.3.3 Sequencing of the full-length *TEP1* genomic sequence

Six primer pairs: VB928/9, 932/33, 936/7, 940/1, 230/1, and 944/5 spanning the whole region (4.8 kb) of the *TEP1* locus were designed to amplify six overlapping fragments (**Fig. 2-2**). The primers were designed using online platforms in http://www.justbio.com/hosted-tools.html and http://www.bioinformatics.org/sms2/. PCR reaction to amplify each fragment, constituted the following components:

0.25 µl (0.2 mM) deoxyribonucleotide triphosphate (dNTP) mix (2 mM each);

2.5 µl (1 ×) 10 × buffer with $MgCl_2$ (1.5 mM);

0.5 µl (5 pmol) each primer (10 pmol/µl);

0.25 µl (0.05 U) Phusion DNA polymerase (2 U/µl); and

21 µl of nuclease free water (NFW) to top up to a total volume of 25 µl.

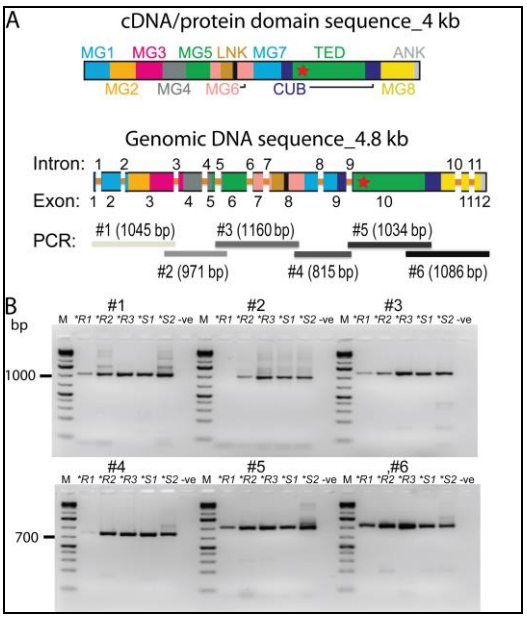


Fig. 2-2. *TEP1* full-length amplification strategy. (A) Schematic representations of coding *TEP1* sequences highlighting the position of exons and introns. Primers for the *TEP1* alleles were designed using the coding regions in order to amplify six overlapping fragments. Primer detail are provided in **Table 2-2**. (B) Expected PCR results of *TEP1* full-length amplification from genomic DNA. Each primer pair was used independently to amplify a specific DNA fragment for each allele from *TEP1* homozygote mosquitoes. The PCR products were purified and cloned for sequencing. Sequence chromatograms were curated in Bioedit and DNA star SeqMan Pro. BLAST searches for each fragment confirmed that all sequences matched the *TEP1* locus.

PCR thermocycling parameters were carried as follows:

98 °C for 30 s;

35 cycles of 98 °C for 15 s, 62 °C for 30 s, and 72 °C for 40 s; and

72 °C for 3 min.

Table 2-2. Primer used for *TEP1* PCR amplification.

Pair	5′-3′ forward primer (VB number)	5′-3′ reverse primer (VB number)	Specificity/ *Taq* polymerase annealing temperature/Applicon (size in bp)	Reference
1	*ATGTGGCAGTTCATA AGGTCAC* (VB928)	*ATCCTGCTGATCCG CATCC* (VB929)	Universal/58°C/exon 1-3#1 (1042-7 bp)	This study
2	*GTCAACGTTCGACG TGCAG* (VB932)	*CGGAATCATCTTTTC TGTTGCGT* (VB933)	Universal/58°C/exon 3-6#2 (993 bp)	This study
3	*GAAGATGTGAATAA GGTAGAAACGG* (VB936)	*TCGGTCTGGTTGGC CACAT* (VB937)	Universal/58°C/exon 5-8#3 (1160 bp)	This study
4	*GTAGTACCGGACAC GACCA* (VB940)	*TCGTAGCTTTGTCGA TTAGATGC* (VB941)	Universal/58°C/exon 8-10#4 (857±2 bp)	This study
5	*TTGGACATCAACAAG AAGGC* (VB230)	*ACTTCAGTTGAACG GTGTAGT* (VB231)	Universal/55°C/exon 9-10#5 (1034±1 bp)	This study
6	*CATACGACCTATCGA TAGCAAC* (VB944)	*GCACTCTGCAGGAC AGTCT* (VB945)	Universal/58°C/exon 10-12#6 (1093-6 bp)	This study
7	*GATGTGGTGAGCAG AATATGG* (VB003)	*ACATCAATTTGCTCC GAGTT* (VB004)	Universal/57°C/Generate primary PCR (892 bp)	Modified from (*80*)
8	*ATCTAATCGACAAAG CTACGAATTT* (VB001)	*CTTCAGTTGAACGGT GTAGTCGTT* (VB002)	Universal/56°C/Generate (*TEP1-TED*) secondary PCR (758 bp)	Modified from (*80*)
9	*TCAACTTGGACATCA ACAAGAAG* (VB229)	*ACATCAATTTGCTCC GAGTT* (VB004)	Universal/53°C/Generate primary PCR (1088±1 bp)	Modified from (*80*, *138*)
10	*TTGGACATCAACAAG AAGGC* (VB230)	*ACTTCAGTTGAACG GTGTAGT* (VB231)	Universal/55°C/Generate secondary PCR (1034±1 bp)	This study
11	*TTGCATGCCATCGG GTCGAAA* (VB221)	*CGGTGAGAAACACG CTACCATT* (VB222)	**R1*-specific/59°C/**R1* detection (148 bp)	Modified from (*80*, *138*)
12	*TTGCATGCCATCGG GTCGAAA* (VB221)	*AACCGTTCGTTTTTA TCAGCATCAATGAA* (VB224)	**R*-specific/59°C/**R* detection (560 bp)	This study
13	*TTGCATGCCATCGG GTCGAAA* (VB221)	*CTATTGGATTCGTTG TGTTCCAGA* (VB228)	**R*-specific/59°C/**R* detection (583 bp)	This study
14	*TTGCATGCCATCGG GTCGAAA* (VB221)	*CTATTTGATTCTTTG TTCTCCAAAACC* (VB232)	**R3*-specific/58°C/**R3* detection (583 bp)	This study
15	*GGTTTGTGGGAGAC TACTGG* (VB223)	*AACCGTTCGTTTTTA TCAGCATCAATGAA* (VB224)	**R2*-specific/59°C/**R2* detection (452 bp)	Modified from (*80*, *138*)
16	*GGTTTGTGGGAGAC TACTGG* (VB223)	*CTATTGGATTCGTTG TGTTCCAGA* (VB228)	**R2*-specific/59°C/**R2* detection (475 bp)	This study
17	*CGGTAAAGTGTGGC ACAAAGAT* (VB225)	*AACCGTTCGTTTTTA TCAGCATCAATGAA* (VB224)	**S2*-specific/58°C/**S2* detection (288 bp)	This study
18	*CGGTAAAGTGTGGC ACAAAGAT* (VB225)	*CTATTGGATTCGTTG TGTTCCAGA* (VB228)	**S2*-specific/58°C/**S2* detection (311 bp)	This study
19	*CGGTAAAGTGTGGC ACAAAGAT* (VB225)	*CAATTTGGTCAGCG CTTTAAGG* (VB227)	**S*-specific/58°C/**S* detection (466 bp)	This study
20	*TCAGTGATAATAATA AAAAAGAACGGTAC* (VB226)	*CAATTTGGTCAGCG CTTTAAGG* (VB227)	**S1*-specific/58°C/**S1* detection (206 bp)	This study

VB stands for Vector Biology, MPI laboratory in Berlin where the thesis work was done. The VB number was assigned to each primer. All the sequenced *TEP1* genomic sequences are available in the NCBI GenBank under accession numbers MF098568 to MF098592 (full-length sequences) and MF035727 to MF035924 (*TEP1-TED* sequences).

PCR fragments were visualized on 1.5% Tris-acetate EDTA (TAE) ethidium-bromide stained agarose gel. The PCR products were cloned into pJet1.2 blunt cloning vector (Thermo Fisher Scientific).

Recombinant plasmids were purified using the miniprep kit (Qiagen) and sent for Sanger sequencing (Eurofins, Germany). Sequence chromatograms were curated manually by visual inspection using Bioedit software *(139)* and Seqman Pro (DNAStar). Sequence fragments amplified from an individual mosquito sample, were further crosschecked by BLAST search in the VectorBase database to confirm the absence of co-amplification with other *TEP* genes. Curated fragments and quality chromatograms were assembled to form a complete *TEP1* full-length contig using Seqman Pro (DNAStar). The full-length sequences are available in the NCBI GenBank under accession numbers MF098568 to MF098592.

2.3.3 *TEP1* genotyping methods

Most genetic markers that are used in the population studies are highly polymorphic and some are neutral i.e. should not be adaptive to the environment (*140*). For example, microsatellites are largely considered neutral and so are under no influence of natural selections because they are located in non-coding regions (*140*). The exceptional polymorphism in *TEP1* locus is characterized by allele-specific amino acid residues within the thioester domain (*TED*) (*5*, *7*, *8*, *76*, *77*).

To explore whether the SNPs coding for these amino acid residues could be harnessed as genetic markers for high throughput *TEP1* genotyping, five SNPs matching restriction sites for *Bam*HI, *Hin*dIII, *Bse*NI and *Nco*I enzymes were chosen and validated using a nested-Polymerase Chain Reaction-restriction fragment length polymorphism (PCR-RFLP) approach. To confirm that the restriction sites for the SNPs were not subject to natural selection (negative or positive) within each allelic subclass, the restriction sites were tested (using selection tools hosted by the Datamonkey server at http://www.datamonkey.org/). Reliably, the restriction sites were only polymorphic between allele groups and fixed per each allele classes or subclasses. In addition, only with exception of a nucleotide (codon **S1105**) for the *Hin*dIII sites in >10% of *TEP1 *S1/S1* in *A. merus* populations in Kenya, all the restriction sites were under no selection (**Table 2-3**), suggesting their suitability as markers for typing all *TEP1* alleles.

Table 2-3. Codons of the SNP genetic markers used in the PCR-RFLP for *TEP1* genotyping, and whether or not the codons are under forces of natural selection.

Type of codon position	**R1*	**R2*	**S1*	**S2*
*Bam*HI SNP sites for **R*	G1019, S1020	G1019, S1020	-	-
*Hin*dIII SNP sites for **S*			G1104, **S1105**, F1106	G1104, S1105, F1106
*Nco*I SNP sites for **S* and **R3*			T843, M844, V845	T843, M844, V845
*Bse*NI SNP sites for **R2*	-	T918, G919	-	-
*Bse*NI SNP sites for **S1*	-	-	Y1063, W1064	-
a) Negatively selected sites	I998	-	Q897, H959, G962, G982, L1036, G1065, T1072, **S1105**	F927, G962, N1068, P1107, G1115
b) Positively selected sites	-	-	AV1004, PA1024	DNR1065

Fig. 2-3A shows schematic representation of *TEP1* genotyping using the PCR-RFLP at the *TED* region. The PCR product is a 758 bp exon that corresponds to positions 2573 to 3390 in reference to *TEP1* full-length coding DNA sequences. Polymorphism at positions 3055-3060 (α7 loop at G1019) carries a *Bam*HI site fixed for **R* alleles that was used to detect the presence of only **R1, *R2* and **R3* alleles, that gives two bands of 399 bp and 365 bp fragments in RFLP digestions. Polymorphism at position 3312-3317 (pre-α12 at G1104) has a *Hin*dIII site in **S* alleles for detecting both **S1* and **S2* alleles by RFLP, resulting in two 656 bp and 102 bp bands. In **S1* alleles, polymorphism at positions 3193-3198 (β-hairpin at Y1063) introduces a *Bse*NI site for detection of the **S1* alleles *via* RFLP, resulting in two fragments of 537 bp and 221 bp.

The **R2* alleles bear a different *Bse*NI site at positions 2757-2762 (pre-α4 at G919) that produces 102 bp and 656 bp RFLP fragments.

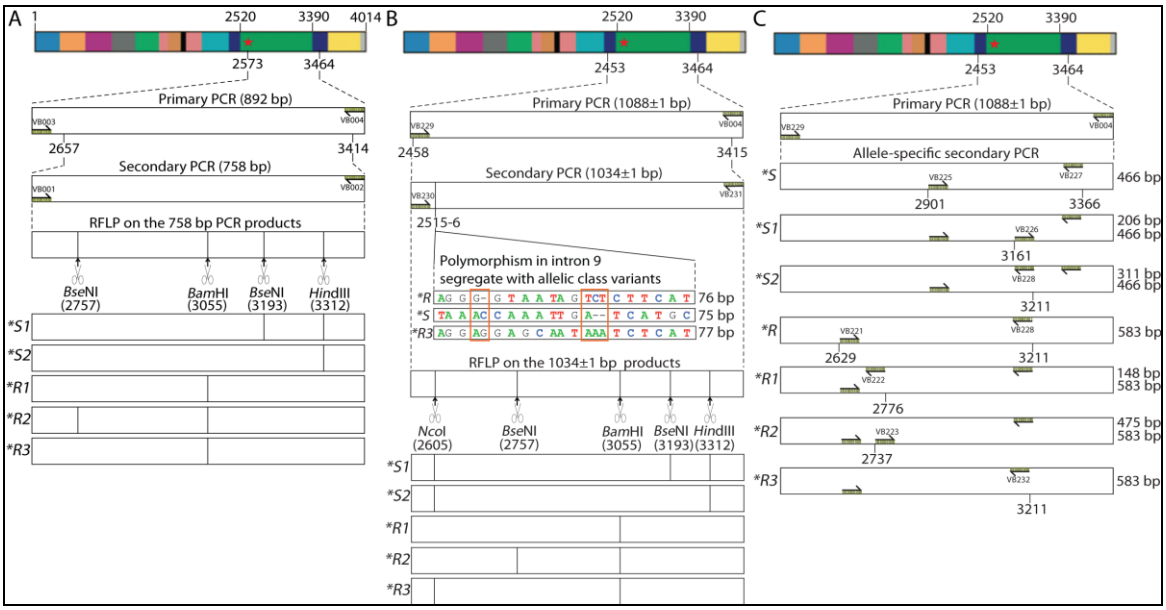


Fig. 2-3. Schematic representation of *TEP1* genotyping methods.
(A) Schematic representation of *TEP1* nested PCR-RFLP genotyping principle (Pompon and Levashina, 2015). It amplifies a final product of 758 bp of the *TED* region using universal primers that can amplify any of the all the four *TEP1* alleles), and the digestions of RFLP on the 758 bp PCRII products. The name and color codes of the primers are indicated on the primers.
(B) Schematic representation of *TEP1* nested PCR-RFLP genotyping method (based on the 1034±1 bp fragment) with the modified primers and the use of the *Nco*I restriction site to complement the *Hin*dIII site. By using the *Nco*I restriction digestion can genotype *TEP1*S1m* (in *Mut6* mutant strain) and *TEP1*R3* alleles.
(C) Schematic representation of *TEP1* PCR-based allele-specific genotyping method.

I used the leg of a single mosquito to extract gDNA for use as a template in primary PCR (PCRI) reaction using primers VB003 and VB004 (**Fig. 2-3**A; **Table 2-2**). The PCRI was done (Fermentas, USA) in total volume of 20 µl containing:

2 µl (0.2 mM) dNTP mix (2 mM each);

2 µl (1×) 10× buffer with 1.5 mM $MgCl_2$;

0.4 µl (0.2 pmol) each primer (10 pmol/µl);

0.4 µl (0.05 U) *Taq* DNA polymerase (5 U/µl);

1 µl (>10 ng/µl) template (gDNA or one mosquito leg); and

14.2 µl of NFW to top up to a total volume of 20 µl.

PCR was carried as follows:

95 °C for 5 min;

20 cycles of 94 °C for 30 s, 57 °C for 30 s, 72 °C for 55 s; and

72 °C for 3 min.

For the secondary or the second nested PCR (PCRII) reaction, 2 µl of primary PCR product was used as a DNA template. Primers VB001 and VB002 (**Table 2-2**) were used in a total PCR reaction (Fermentas, USA) volume of 25 µl containing:

2.5 µl (0.2 mM) dNTP mix (2 mM each);

2.5 µl (1×) 10× buffer with $MgCl_2$ (1.5 mM);

0.4 µl (0.2 pmol) primers (10 pmol/µl);

0.25 µl (0.05 U) *Taq* DNA polymerase (5 U/µl);

2 µl (~10 ng/µl) of primary PCR product; and

16.95 µl of NFW to top up to a total volume of 25 µl.

PCR thermocycling parameters were as follows:

95 °C for 5 min;

40 cycles of 94 °C for 30 s, 56 °C for 30 s, 72 °C for 50 s; and

72 °C for 3 min.

Alternatively, the above PCR amplifications were performed with GO *Taq* kit, which has ready-to-use PCR master mix with a premixed buffer containing dNTP mix, $MgCl_2$ and *Taq* DNA polymerase (Promega, USA).

PCRII products were digested with two sets of restriction enzymes (**Table 2-4**).

Set 1 - a double digestion with *Bam*HI and *Hin*dIII (Thermo Fisher Scientific, Germany) in 37 °C for 1 h in 25 µl reaction volume containing:

2.5 µl (1×) 10 × buffer;

1 µl (10 U) *Bam*HI (10 U/µl);

1 µl (10 U) *Hin*dIII (10 U/µl);

5 µl PCRII product; and

15.5 µl NFW to top the volume to 25 µl.

Set 2 - a single digestion reaction with *Bse*NI (Thermo Fisher Scientific, Germany) at 65 °C for 2 h containing:

2.5 µl (1 ×) 10 × buffer;

1 µl (10 U) *Bse*NI (10 U/µl);

5 µl PCRII product; and

16.5 µl NFW to top the volume to 25 µl.

Digestion products were resolved on 2% agarose gel stained with ethidium bromide run at 10 V/cm length of the gel for 45 min. Patterns of electrophoretic separation were visualized and photographed under ultraviolet (UV) light (**Fig. 2-4**A; **Table 2-4**). Alternatively, the RFLP products were resolved by capillary electrophoresis on a fragment analyzer (Advance Analytical Technical, USA). ProSize software version 2.0

(Advance Analytical Technical, USA). Further analyses on the RFLP fragments were done using an R script (*141*) customized for expected fragments of the PCR-RFLP genotyping (**Table 2-4**).

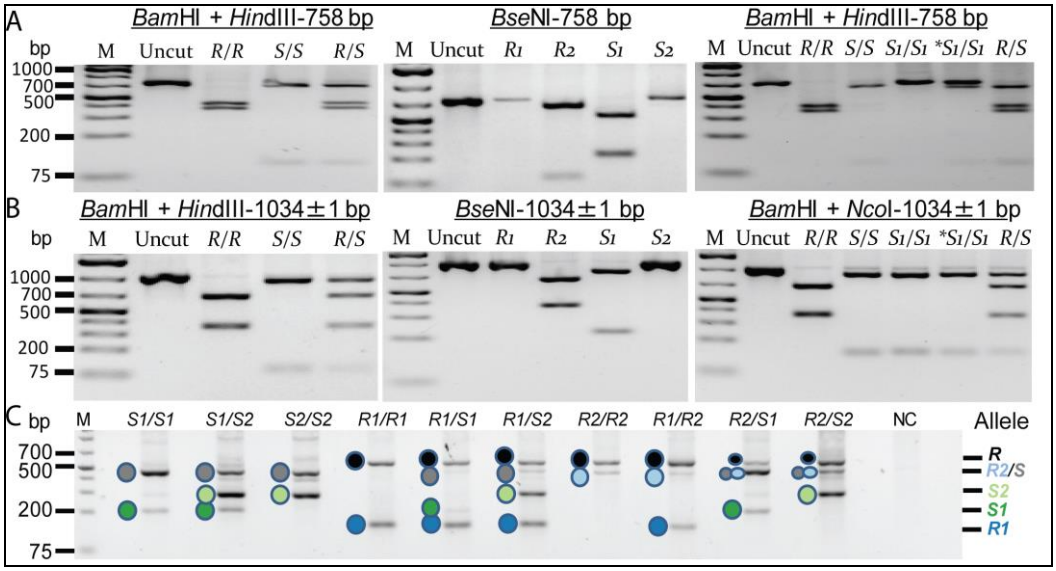


Fig. 2-4. Expected PCR results for *TEP1* genotyping of **R1, *R2, *S1* and **S2* alleles.
(A) Fragments of expected sizes produced as a result of digest of the amplified region of *TEP1* were used for allele determination. Digest with *Bam*HI identifies **R* alleles at the α7 loop, while digest with *Hin*dIII only cuts **S* amplicons at pre-α12 loop. Digestion with *Bse*NI cuts only **R2* and **S1* amplicons at the pre-α4 and β-hairpin loops respectively. The last panel shows representative results of *TEP1*S1/S1 A. merus* from Kenya.
(B) Results of *TEP1* nested PCR-RFLP genotyping using modified primers and the *Nco*I restriction site at α0 loop M844 instead of the *Hin*dIII digestion. Sequencing and the *Nco*I restriction pattern confirmed the identity of *TEP1*S1/S1* samples. The same method was used to genotype **R3* allele.
(C) *TEP1* PCR-based allele-specific genotyping strategy was used to confirm the results of PCR-RFLP if necessary. The results are from the multiplex PCR with all 7 primers, NC is a negative control. The name and color codes on the gel correspond to each allele. Note that the best results were achieved by using primer pairs for 1 to 2 alleles at a time.

To identify *A. merus TEP1*S1* alleles that lacked *Hin*dIII site, a 1034±1 bp PCRII fragments were amplified and subjected to a double *Bam*HI and *Nco*I digest (**Fig. 2-3**B; **Fig. 2-4**B; **Table 2-5**). Similar to **S*, the **R3* allele also had an *Nco*I site at the same positions 2605-2610 (α0 loop M844).

For samples that could not be genotyped by the PCR-RFLP genotyping methods, I designed a PCR-based genotyping (**Fig. 2-3**A; **Table 2-6**). To this end, published *TEP1* allele-specific PCR primers were modified (*7*, *8*) to allow genotyping of all the known alleles.

Table 2-4. Expected RFLP fragment sizes (bp) of *TEP1* genotypes resulting from a digest of the 758-bp *TEP1* amplicon.

	Restriction enzyme and expected fragment sizes	
TEP1 genotype	*Bam*HI+*Hin*dIII	*Bse*NI
S1/S1	656, 102	537, 221
S1/S2	656, 102	758, 537, 221
S2/S2	656, 102	758
R1/R1	399, 359	758
R1/S1	399, 359, 656, 102	758, 537, 221
R1/S2	399, 359, 656, 102	758
R2/R2	399, 359	657, 101
R2/S1	399, 359, 656, 102	537, 221, 657, 101
R2/S2	399, 359, 656, 102	758, 657, 101
R2/R1	399, 359	758, 657, 101
R3/R3	399, 359	758
R3/S1	399, 359, 656, 102	758, 537, 221
R3/S2	399, 359, 656, 102	758
R3/R1	399, 359	758
R3/R2	399, 359	758, 657, 101

Table 2-5. Expected RFLP fragment sizes (bp) of *TEP1* genotypes resulting from a digest of the 1034±1 bp *TEP1* amplicon.

	Restriction enzyme and expected fragment sizes		
TEP1 genotype	*Bam*HI+*Hin*dIII	*Bam*HI+*Nco*I	*Bse*NI
S1/S1	930, 103	146, 887	811, 222
S2/S1	930, 103	146, 887	1033, 811, 222
S2/S2	930, 103	146, 887	1033
R1/R1	674, 360	674, 360	1034
R1/S1	674, 360, 930, 103	674, 360, 146, 887	1034, 811, 222
R1/S2	674, 360, 930, 103	674, 360, 146, 887	1034*, 1033*
R2/R2	674, 360	674, 360	376, 658
R2/S1	674, 360, 930, 103	674, 360, 146, 887	811, 222, 376, 658
R2/S2	674, 360, 930, 103	674, 360, 146, 887	1033, 376, 658
R2/R1	674, 360	674, 360	1034, 376, 658
R3/R3	675, 360	527, 360, 148	1035
R3/S1	675, 360, 930, 103	887, 527, 360, 146*, 148*	1035, 811, 222
R3/S2	675, 360, 930, 103	887, 527, 360, 146*, 148*	1035*, 1033*
R3/R1	675*, 674*, 360	674, 527, 360, 148	1035*, 1034*
R3/R2	675*, 674*, 360	674, 527, 360, 148	1035, 376, 658

*Fragments have close size ranges and would appear as one overlapped fragment.

Table 2-6. Expected fragment sizes (bp) from *TEP1* PCR-based genotyping.

	Allele-specificity of the primer and expected fragment sizes						
TEP1 genotype	**R*	**R1*	**R2*	**R3*	**S*	**S1*	**S2*
R1/R1	583[$]	148	-	-	-	-	-
R1/R2	583[$]	148	475[$]	-	-	-	-
R1/R3	583[$]	148	-	583	-	-	-
R1/S1	583[$]	148	-	-	466	206	-
R1/S2	583[$]	148	-	-	466	-	311[$]
R2/R2	583[$]	-	475[$]	-	-	-	-
R2/R3	583[$]	-	475[$]	583	-	-	-
R2/S1	583[$]	-	475[$]	-	466	206	-
R2/S2	583[$]	-	475[$]	-	466	-	311[$]
R3/R3	583[$]	-	-	583	-	-	-
R3/S1	583[$]	-	-	583	466	206	-
R3/S2	583[$]	-	-	583	466	-	311[$]
S1/S1	-	-	-	-	466	206	-
S1/S2	-	-	-	-	466	206	311[$]
S2/S2	-	-	-	-	466	-	311[$]

[$]When VB228 (**S2* - and **R*- specific) primer is used in PCR instead of VB224.

2.3.4 *TEP1* sequencing and sequence analyses

The *TEP1-TED* 758 bp amplicons were cloned either into pGEM-T Easy (Promega, USA) vector or pjet1.2 vector (Thermo Fischer Scientific, USA). The clones were sequenced from both orientations (Eurofins, Germany). Sequence chromatograms were curated and aligned using the Bioedit version (*139*). Various sequence manipulation free-hosted tools in justbio.com and http://www.bioinformatics.org/sms2/ were used for *in silico* translation of the sequences into amino acids as well as digestions with *Bam*HI, *Hin*dIII and *Bse*NI restriction enzymes. DnaSP version 5.0 (*142*, *143*) was used to analyze DNA polymorphism and neutrality tests such as Tajima's D tests, Fu & Li's D to infer selection pressures. *TEP1* DNA sequences were collapsed into network of allele haplotypes using TCS1.21 (*144*). The TCS1.21 software uses a statistical parsimony to infer gene genealogies. It generates genealogy network based on mutational steps that separate or connect the haplotypes and is able to detect back mutations.

Both *TEP1-TED* nucleotide and amino acid sequence alignments were used for phylogenetic analyses. Jmodeltest (*145*) was used to choose the most optimal nucleotide substitution model for the DNA sequence data sets (*145*, *146*). The Jmodeltest analyses suggested parameters of K80 (*147*) with invariable sites and gamma rate of evolution. DNA sequences were translated into amino acid sequences as done above. Distance-

based neighbor-joining and maximum likelihood-based phylogenetic algorithms on the sequences were performed using MEGA6 (*148*), PhyML 3.0 (*149*) and BEAST (*150*) and produced similar tree topologies. The reliability of these analyses was evaluated by bootstrap tests with 1000 replications. All the positions with gaps and missing data were eliminated from the dataset. Phylogenetic trees were annotated in MEGA6 (*148*), and adobe illustrator CS5 and Photoshop CS5 (Adobe Systems, California USA). Full-length multiple sequence alignments presented in the Appendix **5** and Appendix **6** were done with Kalign in http://www.ebi.ac.uk/Tools/msa/ and Boxshade in http://www.ch.embnet.org/software/BOX_form.html online platforms.

2.3.5 Statistical analyses

The data on alleles and genotypes were analyzed in R. version 3.1.3 (2015) (*141*). Customized R scripts and R-packages: r genetic package and standard graphical packages such as plyr, ggplot2, reshape2, gridExtra, cowplot for data manipulation and plotting of graphs (Appendix **1**D). Chi-Square tests were used to analyse for deviations from expectations of the Hardy Weinberg Equilibrium. Part of the R scripts is provided in the Appendix **2** and Appendix **3**.

2.4 Results

2.4.1 Overview of study countries and *A. gambiae s.l.* samples

A total of 1556 of *Anopheles s.l.* mosquitoes were sampled from Mali, Burkina Faso both in West Africa, Cameroon in Central Africa, and Kenya in East Africa, and were molecularly identified to species level (**Fig. 2-5**A).

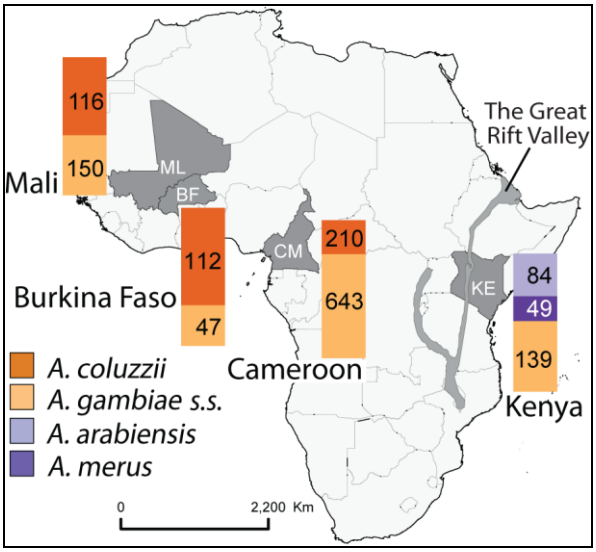


Fig. 2-5. Composition of *A. gambiae s.l.* samples from sub-Saharan African countries.
The samples that were collected in Mali, Burkina Faso, and Cameroon were *A. coluzii* and *A. gambiae s.s.*. In Kenya, *A. arabiensis, A. gambiae s.s.* and *A. merus* were collected. The stacked bar charts show species distribution and the numbers in them represent the number of samples collected in each country per species.

This constituted the following mosquito species: *A. coluzzii* (*n* = 116) and *A. gambiae s.s.* (n = 150) in Mali; *A. coluzzii* (*n* = 112) and *A. gambiae s.s.* (*n* = 47) in Burkina Faso; *A. coluzzii* (*n* = 210) and *A. gambiae s.s.* (*n* = 643) in Cameroon; and *A. arabiensis* (*n* = 84)*, A. gambiae s.s.* (*n* = 139) and *A. merus* (*n* = 49) in Kenya.

2.4.2 *TED* region resolves natural *TEP1* variation into distinct allelic subclasses

To characterize *TEP1* genetic variation in the field-sampled malaria vectors, first a robust high-throughput *TEP1* genotyping approach was developed. To do this, mosquito samples from Cameroon (*n* = 21) and Mali (*n* = 2) were used to amplify and sequence full-length *TEP1* genomic sequences, representing **R1, **R2, **S1* and **S2* allelic subclasses.

The sequences were curated to correct ambiguous base calls, and subsequently, all the 11 introns were trimmed off leaving only 4014 bp DNA coding sequences for analyses. Allele-specific SNPs located at the *TED* region were used to categorize *TEP1* sequences into **R1, **R2, **S1* and **S2* allelic subclasses. The full-length *TEP1* sequences of field collected samples were aligned together with those from four laboratory strains; L3-5 (**R1*), 4Arr (**R2),* G3S3 (**S1*) and 4Arr (**S2)* bearing **R1, **R2, **S1* and **S2* alleles, respectively.

To characterize distribution of synonymous (dS) and non-synonymous (dN) mutations in *TEP1* full-length sequences, cumulative behaviour of dS and dN substitutions was calculated. Results reveal unique differences in the substitution and selective pressure regimes (**Fig. 2-6**). Specifically, **R1* alleles had excess of dN compared to dS substitutions (0 - 10), **R2* alleles had the most balanced dS and dN (0 - 3), **S1* alleles featured excess of dS compared to dN substitutions (0 - 20) with the widest allele diversity, and **S2* alleles also had excess of dS compared to dN substitutions (0 - 5) but it remained fairly constant over the gene locus.

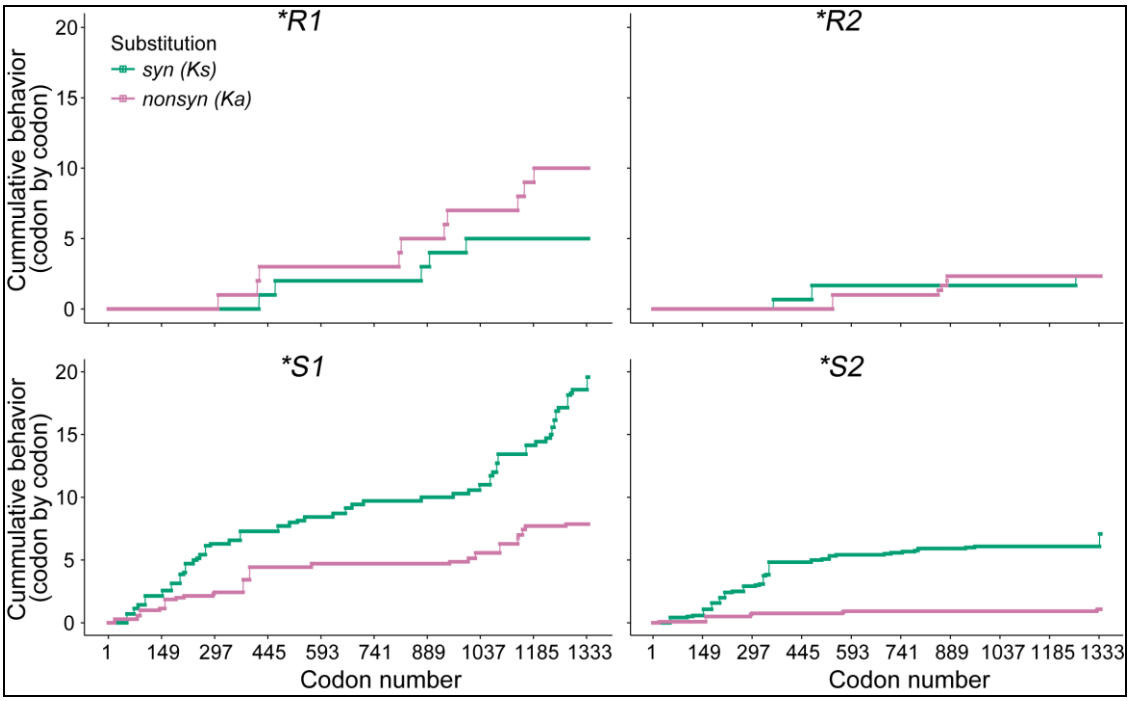


Fig. 2-6. Codon diversity and variability in behavior of *dN* and *dS* substitutions reveals allele-specific selective forces acting on *TEP1* locus. Full-length coding DNA sequences **R1* (*n* = 2), **R2* (*n* = 4), **S1* (*n* = 8) and **S2* (*n* = 13) were used to calculate codon-based cumulative synonymous (*dS*) and non-synonymous (*dN*) substitutions using SNAP online platform (*151*, *152*) . Codon data were exported to R for plotting (*141*).

Analysis of DNA polymorphism was carried out using the DnaSP version 5 (*142*, *143*) within and between allele classes. No significant differences were observed with the neutrality tests; Tajima's D, Fu and Li's D and Fu and Li's F statistics (**Table 2-7**).

Table 2-7. Neutrality test on *TEP1* full-length coding sequences.

Statistics	Allele				
	All alleles	**S1*	**S2*	**R1*	**R2*
Tajima's D	-0.02	0.44	-1.00	NA	-0.79
Fu and Li's D	-0.07	0.65	-1.45	NA	-0.79
Fu and Li's F	-0.07	0.67	-1.52	NA	-0.84
Sequences No.	26	8	12	2	4
Haplotypes No.	20	7	7	2	4
Haplotype diversity	0.98	0.96	0.89	1.000	1.000
Variable sites No.	367	65	35	15	78
Total mutations	375	67	35	15	78
Nucleotide diversity/site	0.024	0.007	0.002	0.004	0.010

NA = Not applicable

However, sliding window analyses revealed highest levels of nucleotide diversity (Pi) scores at the *TED* region (0.01 - 0.09) as compared to other domains (0 - 0.03) (**Fig. 2-7**A).

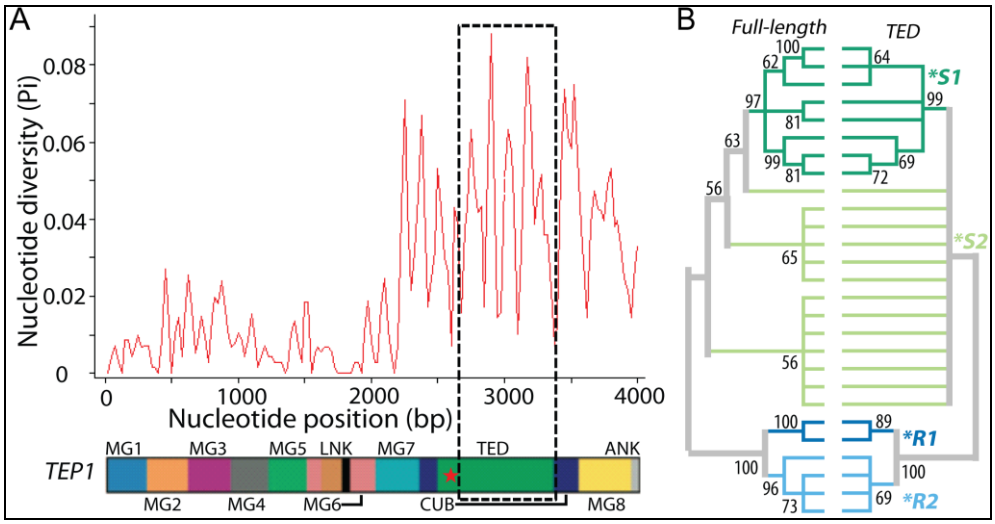


Fig. 2-7. Genetic diversity of *TEP1* locus.
(A) Sliding window plot of nucleotide diversity of full-length coding sequences over TEP1 protein coding sequence. In total, 27 sequences as used in Fig. 2-6; Cameroon (n = 21), Mali (n = 2) and laboratory strains (n = 4) representing four allelic subclasses were used. Highlighted in dotted window is the 758 bp *TED* region that was used to develop high throughput *TEP1* genotyping approach.
(B) Comparative topology of two neighbor-joining (NJ) trees (50% cut-off values, 1000 bootstrap replicates, number of differences) for 4014 bp full-length (left) and 758 bp *TED* (right) sequences.

To validate whether variability at the *TED* region mirrors that of the full-length *TEP1*, I performed neighbor joining sequence comparison of *TEP1* full-length and *TED* sequences (highlighted with a dotted window in **Fig. 2-7**A). Phylogenetic trees based on the alignment of full-length (left) and TED region (right) showed similar topologies (**Fig. 2-7**B, left panel; **Fig. 2-7**B, right panel), although full-length sequences contained more information for fine resolution of the **S2* allele cluster. Therefore, this study entirely focused on the *TED* region and used it to develop a high throughput nested *TEP1* PCR-RFLP genotyping approach. Five genetic markers corresponding to *Bam*HI, *Hin*dIII, *Nco*I and *Bse*NI restriction sites were identified and validated in the PCR-RFLP method for typing all the alleles used in this study (see Materials and Methods).

2.4.3 *TEP1* genotypes across Africa

To identify distribution of *TEP1* genotypes in the selected mosquito populations, a total of 1556 mosquitoes were genotyped. Overall, 11 (five homozygous and six heterozygous) *TEP1* genotypes with varied frequencies were identified across Africa (**Fig. 2-8**; **Fig. 2-9**; **Fig. 2-10**A-C). An R-script used to generate **Fig. 2-9** and **Fig. 2-10** is provided in the Appendix **2**B-C.

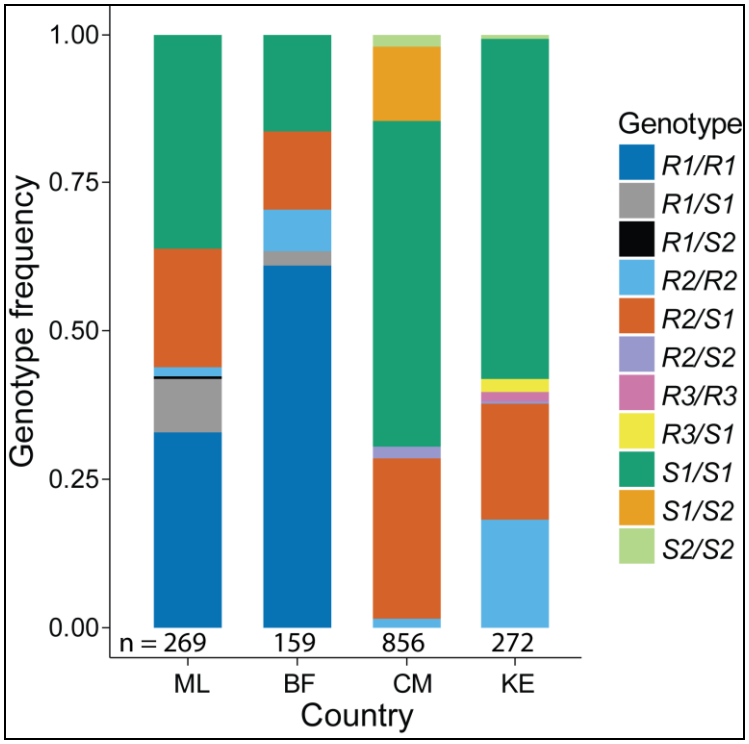


Fig. 2-8. Global distribution of *TEP1* genotypes in Africa.
Overview of *TEP1* genotypes per country across Africa. The numbers below each column represent the number of genotyped individuals. Abbreviations of countries are as follows: ML, Mali; BF, Burkina Faso; CM, Cameroon; and KE, Kenya.

*TEP1*S1/S1* genotype was the most common genotype present in all the countries. *TEP1*R2/S1* and **R2/R2* were also found in all countries but at lower frequencies. The **R1/R1* genotype was only found in Mali and Burkina Faso. In Mali, Cameroon and Kenya, **S2/S2* genotypes were detected albeit at low frequencies. Interestingly, high number of **S1/S2* genotypes was observed in Cameroon. Interestingly, a new **R*-type allele, hereafter named **R3*, was found in *A. merus* populations along the coastal Kenya (**Fig. 2-8**).

Comparison between the observed and the expected genotypes using the data set for all the countries showed departures from the HWE (**Table 2-8**; **Table 2-9**; Appendix **3**; Appendix **4**). To quantify these deviations, population *F* statistics for subpopulations and global population i.e. all combined mosquito subpopulations were calculated using Wright´s *F* statistics (see the R-script in Appendix **3** and the statistical output in Appendix **4**). Variation was observed in inbreeding coefficients (F_S) and/or homozygosities (F_S <0) between species and sampling sites. *A. gambiae s.s.* in Mali (F_S = -0.1) and in some Cameroonian sites (F_S = -0.06) has deficiency in homozygotes suggesting that they were outbred. In contrast, *A. gambiae s.s.* populations in Burkina

Faso (F_S = 0.27) and Kenya (F_S > 0) showed excess homozygosity or inbreeding. Similarly, in Kenya, *A. arabiensis* populations had high inbreeding coefficient (F_S = 0.65, -0.58, 0.3).

Table 2-8. Inbreeding coefficient (*Fs*).

Country	Site	Species	*Hom*	HET_{Obs}	HET_{Exp}	*Fs*	χ^2_{cal} Statistic
ML	NK	*A. col*	10	0.19	0.27	0.30	$df = 3, \chi^2_{cal} = 10.07$*
ML	NK	*A. gam*	-5.3	0.37	0.34	-0.10	$df = 3, \chi^2_{cal} = 2.09$
BF	SM	*A. col*	-2.7	0.5	0.49	-0.03	$df = 1, \chi^2_{cal} = 0.01$
BF	SM	*A. gam*	24.3	0.34	0.47	0.27	$df = 3, \chi^2_{cal} = 4.51$
CM	MV	*A. col*	2.7	0.48	0.49	0.03	$df = 3, \chi^2_{cal} = 0.90$
CM	MV	*A. gam*	3.1	0.44	0.46	0.04	$df = 3, \chi^2_{cal} = 9.46$*
CM	NS	*A. col*	-0.46	0.50	0.50	0.00	$df = 3, \chi^2_{cal} = 0.53$
CM	NS	*A. gam*	-3.7	0.39	0.36	-0.06	$df = 3, \chi^2_{cal} = 48.69$*
CM	MF+ND+NM	*A. col*	-12.0	0.5	0.43	-0.16	$df = 3, \chi^2_{cal} = 1.88$
CM	MF+ND+NM	*A. gam*	-3.5	0.38	0.35	-0.06	$df = 3, \chi^2_{cal} = 31.52$*
KE	AH	*A. ara*	0.6	0.17	0.48	0.65	$df = 1, \chi^2_{cal} = 25.57$*
KE	AH	*A. gam*	0.82	0.08	0.50	0.83	$df = 1, \chi^2_{cal} = 8.31$*
KE	BT	*A. gam*	0.07	0.23	0.28	0.19	$df = 3, \chi^2_{cal} = 7.62$
KE	KK	*A. ara*	0.53	0.20	0.48	0.58	$df = 1, \chi^2_{cal} = 4.31$*
KE	KK	*A. gam*	0.15	0.07	0.20	0.63	$df = 1, \chi^2_{cal} = 10.55$*
KE	MD	*A. ara*	20.7	0.29	0.41	0.3	$df = 1, \chi^2_{cal} = 0.63$
KE	MD	*A. mer*	32.2	0.34	0.50	0.32	$df = 6, \chi^2_{cal} = 56.46$*

Hom = Homozygosity; HET_{Obs} = Observed heterozgosity; HET_{Exp} = Expected heterozgosity;
Fs = Local inbreeding coefficient; Asterisk (*) indicates χ^2 significant deviation from the HWE at $p<0.05$.

2.4.4 Species-specific distribution of *TEP1* genotypes

Does *TEP1* genotype distribution vary between species? **Fig. 2-9** summarizes distribution of *TEP1* genotypes in the vector species per country samples in this study. Interestingly, moderate frequency of **R1/S1* genotypes were detected in *A. coluzzii* from Mali, whereas in Burkina Faso and in *A. gambiae s.s.* from Mali these genotypes were very rare. Strikingly, we confirm that the divergence between sympatric populations of *A. coluzzii* and *A. gambiae s.s.* is higher in Mali (F_{ST} = 0.47) than in Burkina Faso (F_{ST} = *0.003*) (**Table 2-9**; Appendix **4**), suggesting genetic differentiation. I propose that degree of segregation in sympatric species at the *TEP1* locus is dependent on local breeding habitats.

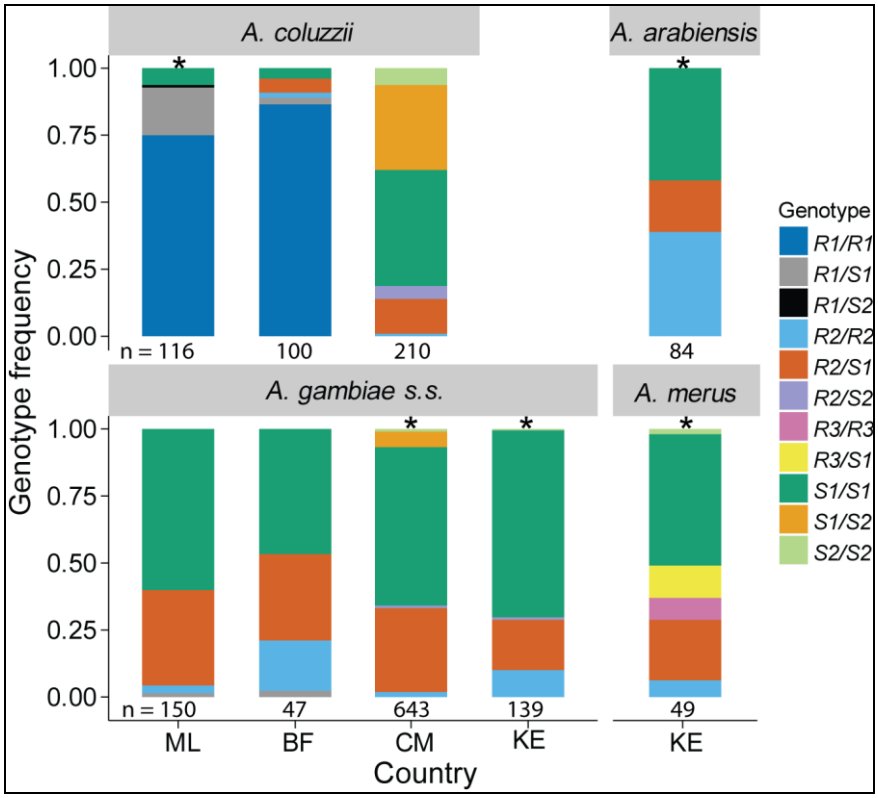


Fig. 2-9. Global view of mosquito vector population species and *TEP1* genotypes. Overview of *TEP1* genotypes per species in Mali (ML), Burkina Faso (BF), Cameroon (CM), Kenya (KE). The numbers below each column indicate the number of genotyped individuals. Asterisk (*) indicates significant deviation from the HWE expection.

Table 2-9. Population Wright´s *F*-statistics in sympatric mosquito populations.

Country	Site	Species	*Hi*	*Hs*	*Ht*	F_{IS}	F_{IT}	F_{ST}
ML	NK	*Col-gam*	0.29	0.31	0.59	0.05	0.50	0.47
BF	SM	*Col-gam*	0.37	0.47	0.47	0.21	0.21	0.003
CM	MV	*Col-gam*	0.47	0.46	0.49	-0.02	0.04	0.06
CM	NS	*Col-gam*	0.42	0.36	0.42	-0.16	-0.01	0.13
CM	MF+ND+NM	*Col-gam*	0.38	0.35	0.36	-0.08	-0.07	0.01
KE	AH	*Ara-gam*	0.15	0.48	0.49	0.68	0.69	0.01
KE	KK	*Ara-gam*	0.09	0.24	0.26	0.61	0.64	0.08
KE	MD	*Ara-mer*	0.33	0.49	0.55	0.32	0.12	0.40

Hi = Observed heterozygosity; *Hs* = Sum of expected heterozygosity within one subpopulation; *Ht* = Sum of expected heterozygosity in all populations; *Fis* = Inbreeding coefficient; F_{ST} = Wright´s standard variance (population differentiation index) between demes; and F_{IT} = Deviation from the expected HWE proportions.

Higher numbers of *TEP1*S1/S2* mosquitoes were identified for *A. coluzzii* in Cameroon (**Fig. 2-9**). In addition, **S2/S2* homozygotes were predominantly found in *A. coluzzii* in Cameroon, whereas this genotype was very rare in other countries and species.

In Kenya, *TEP1***R3/R3* and **R3/S1* genotypes were only found in *A. merus*. *A. arabiensis* species were enriched for **R2/R2* and **S1/S1* homozygotes with proportionally low number of **R2/S1* heterozygotes. Interestingly, similar enrichment of **R2/R2* and **R2/S1* genotypes was observed in *A. merus*, whereas very low frequencies of **R2/R2* and **R2*/*S2* heterozygotes were detected in Cameroon. In Kenya and to the least extent in Mali, the **S2/S2* genotypes appeared at low frequencies.

2.4.5 Local-specific biotope factors determine *TEP1* genotype distribution

How are *TEP1* genotypes stratified at the species and habitat level across Africa? In Mali, *A. coluzzii* and *A. gambiae s.s.* populations breed in sympatry (**Fig. 2-10**A). We detected *A. gambiae s.s.* **R1/S1* individuals as well as low number (1%) of *A. coluzzii* / *A. gambiae s.s.* hybrids with **R1/R1* (n = 2) and **R1/S1* (n = 1) genotypes (**Fig. 2-10**A) but no hybrids were found in Burkina Faso.

In Cameroon, sympatric *A. coluzzii* and *A. gambiae s.s.* mosquitoes from the five different districts (Mfou, Mvan, Nkolondom, Nkolkoumou and Nkolbisson) were collected (**Fig. 2-10**B). Across all the Cameroon sampling sites, an average ratio of **S1/S2* to **R2/S1* individuals in sympatric *A. coluzzii* and *A. gambiae s.s.* populations was 2:1. In addition, this ratio was switched in *A. gambiae s.s.* to *A. coluzzii* species to 1:2. Further, the Cameroon-restricted **S1/S2* genotypes (~30%) showed a negative correlation with **R2/S1* genotypes (>15%). The few *A. coluzzii* / *A. gambiae s.s.* hybrids in Cameroon (1.5%) featured **R2/S1* (n = 4) and **S1/S1* (n = 2) genotypes (**Fig. 2-10**C). To calculate the inbreeding coefficient and F-statistics, Cameroon's sympatric populations in Mfou, Nkolondom and Nkolkoumou were merged since proportions of *A. gambiae s.s.* were similar (>90%), compared with Mvan (25%) and Nkolbisson (69%). In breeding coefficients (Fs >0.19, F_{ST}>0.3) were higher in Kenya than those in Mali, Burkina Faso and Cameroon (**Table 2-8**; **Table 2-9**; **Fig. 2-10**). In addition, the Wright´s fixation indices (**Table 2-9**) showed marked variation in species pairwise comparisons across Africa suggesting existence of population structure at the *TEP1* locus. These analyses were not done for Kilifi and Kwale sites due to low sample sizes.

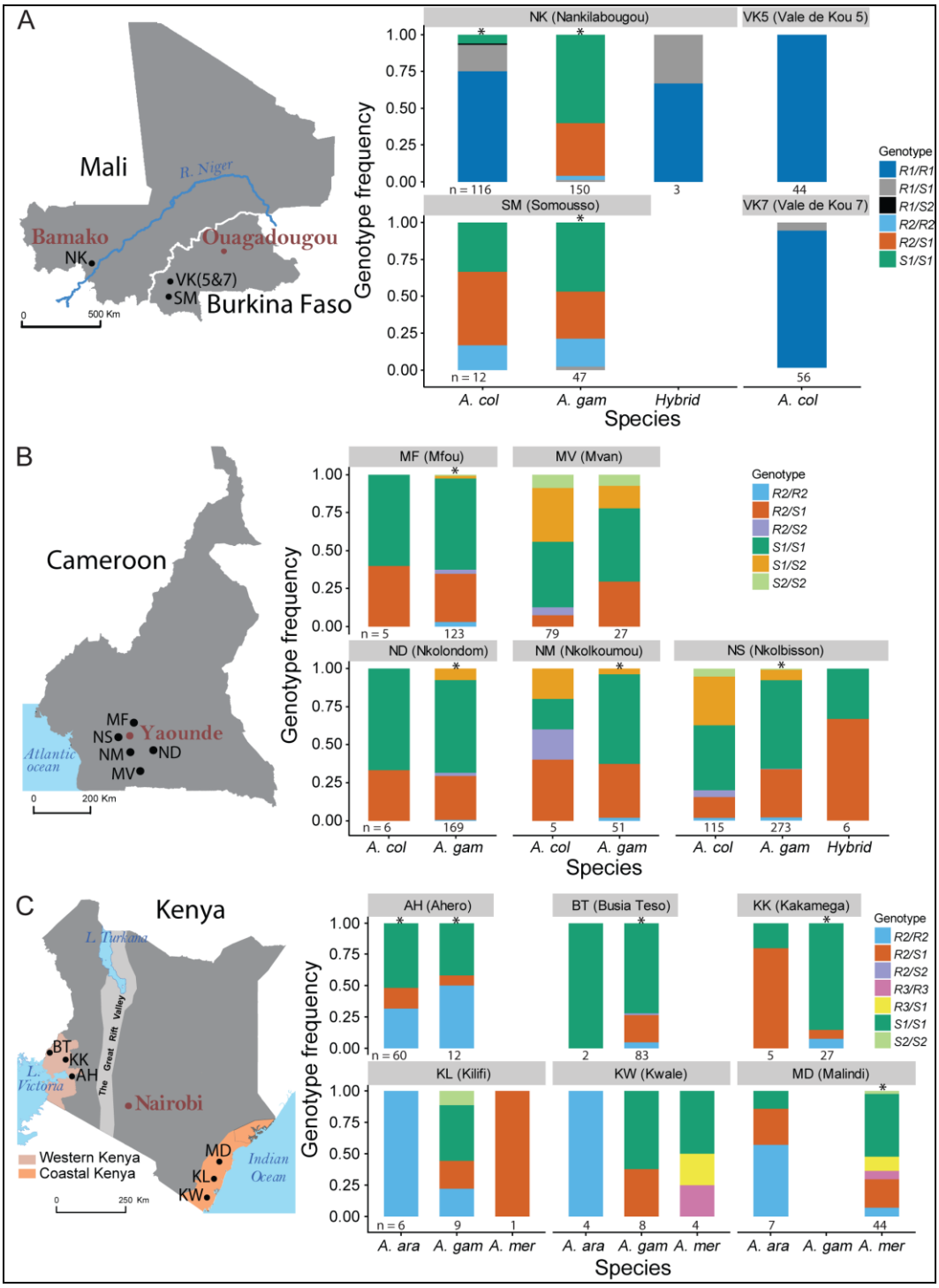


Fig. 2-10. Sampling sites and distribution of *TEP1* genotypes per species per site.
(A) Sampling sites and distribution of *TEP1* genotypes in Mali and Burkina Faso. Abbreviations of mosquito vector species are as follows: *A. col, A coluzzii* and *A. gam, A. gambiae s.s.*. *Hybrid* refers to rare offspring resulting from natural crossbreeding between *A. coluzzii* and *A. gambiae s.s.* species. The numbers shown below each stacked bar chart in A-C, represent the sample sizes of the genotyped individuals per species per site.
(B) Sampling sites and distribution of *TEP1* genotypes in Cameroon. Abbreviation of mosquito vector species *A. col* and *A. gam* are as in A.
(C) Sampling sites and distribution of *TEP1* genotypes in Kenya: Abbreviation of mosquito vector species *A. ara, A. arabiensis; A. gam, A. gambiae s.s.* and *A. mer, A. merus.* Asterisk (*) indicates significant deviation from HWE tests using a conservative χ^2 that corrects for small size sample numbers. See page 23 and the statistical output in Appendix **5**.

2.4.6 Allelic frequencies and inference of genetic relationship

Allelic frequencies were calculated for the merged sites from the *TEP1* genotype data (**Fig. 2-11**). An R-script used to generate this figure is provided in Appendix **3**D. Some alleles were either completely absent or very rare in certain species and in certain sampling locations (**Fig. 2-10**A-C; **Fig. 2-11**). *TEP1*S2* was mostly common in Cameroon, less abundant in Mali and Kenya but absent in Burkina Faso. Significant deviation from the HWE was observed for **R2, *S1* and **S2* alleles in sympatric *A. coluzzii* and *A. gambiae s.s.* population in Cameroon and in *A. coluzzii* populations in Mali. Significant deviations from the HWE expectations were also evident among populations of *A. merus* and *A. gambiae s.s.* in coastal Kenya, and *A. gambiae s.s.* and *A. arabiensis* in western Kenya (**Fig. 2-11**).

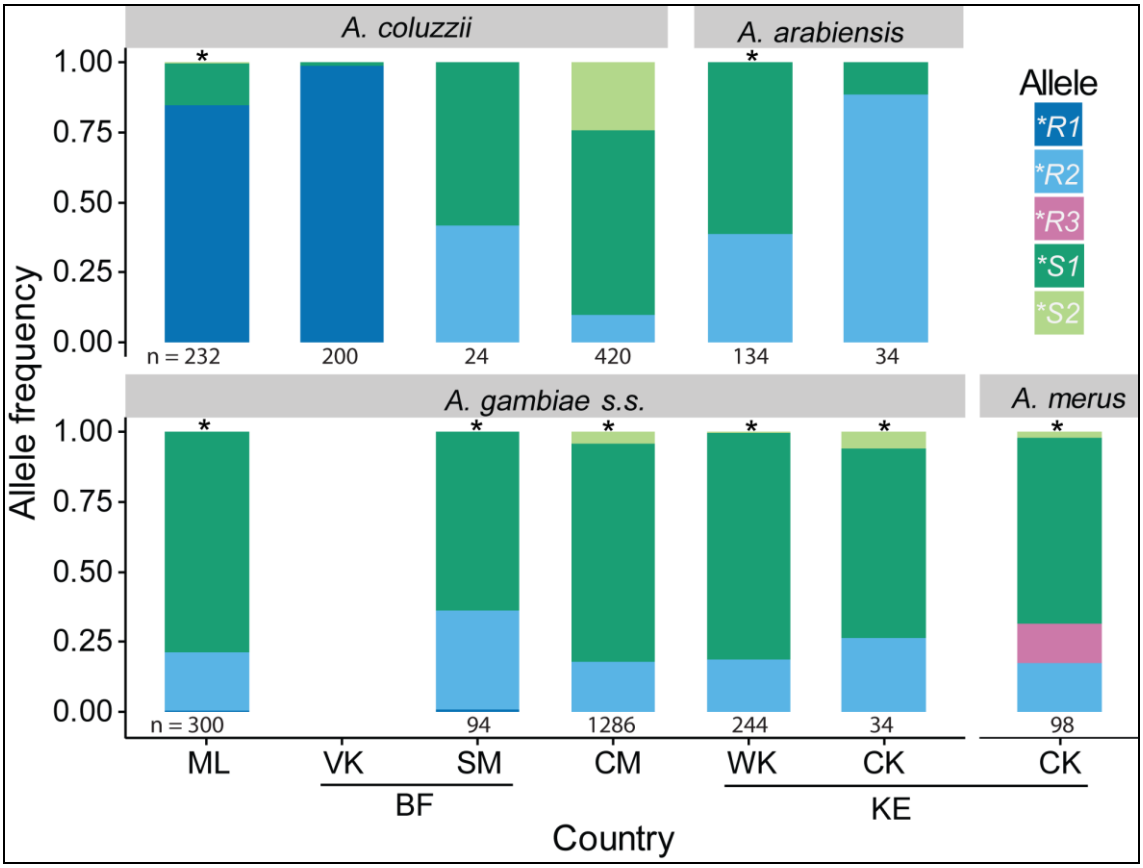


Fig. 2-11. Global *TEP1* allele frequencies across Africa.
Overview of distribution of allele frequencies across African countries. Abbreviations of sampling locations and sites are as follows, ML, Mali; VK5, Vale de kou 5; VK7, Vale de kou 7; SM, Somousso in Burkina Faso; CM, Cameroon (MF, Mfou; MV, Mvan; ND, Nkolondom; NM, Nkolkoumou and NS, Nkolbisson were merged); WK, western Kenya (AH, Ahero; KK, Kakamega and BT, Busia Teso were merged), and CK, coastal Kenya (KW, Kwale; KL, Kilifi and MD, Malindi were merged). The numbers along the x-axes represent the number of sampled alleles. Asterisk (*) indicates significant deviation from the HWE expectations.

While the **R2* and the **S1*, and to a lesser extent the **S2* alleles were conserved across Africa, the private **R1* in *A. coluzzii* and **R3* in *A. merus* were not only the most diverse, but also showed geographical- and species-specific patterns. In overall, **S1* allele appears to be the most compatible allele with all the others as confirmed by the occurrence of the **R1/S1, *R2/S1* and **S1/S2* homozygotes. Therefore, our results uncovered marked differences in the distribution of *TEP1* alleles and/or genotypes between and within the sampling sites as well as between the species.

2.4.7 Sequence analyses

To interrogate the genetic relationship and diversity of the alleles, the *TED* region of 195 (including 3 more from our G3 laboratory strains, to make 198 in total) representative alleles from different species across all the surveyed countries was sequenced and subjected to phylogenetic analyses using various amino acid substitution models, and all showed similar allele clustering (**Fig. 2-12**). The 198 *TED* nucleotide sequences were deposited in the NCBI GenBank under the accession numbers MF035727 to MF035924. A few of published sequences representing different *TEP1* alleles from field-caught (*6*) and laboratory mosquitoes were included in the analyses.

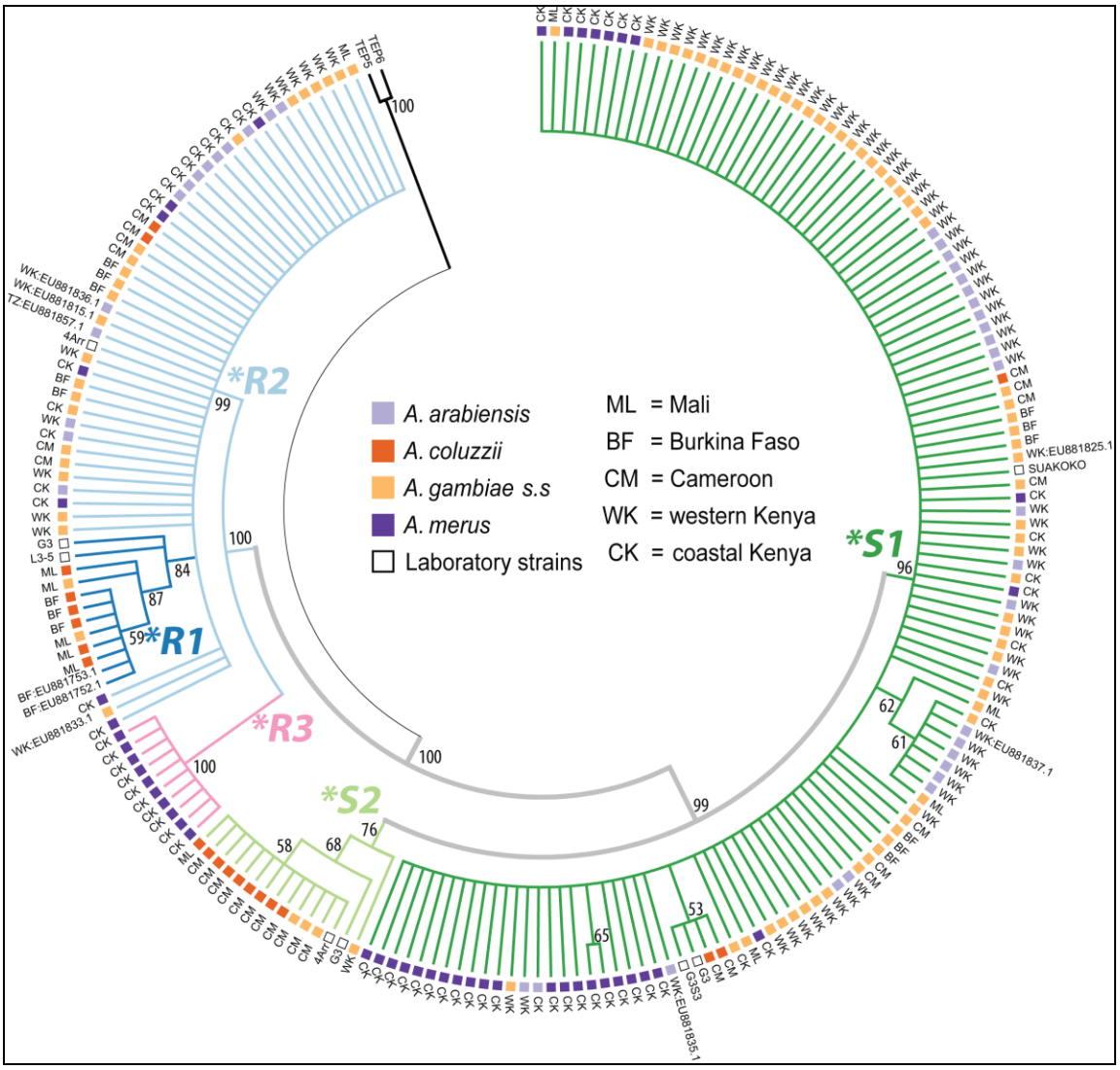


Fig. 2-12. Geodiversity of surveyed species stratified by *TEP1* alleles across Africa**.**
A. gambiae s.l sampled across four geographically diverse sub-Saharan countries. *TEP5* and *TEP6* were used as outgroup taxa. The 198 *TED* nucleotide sequences from our data set were deposited in the NCBI GenBank under accession numbers MF035727 to MF035924. Sequences of other published *TEP1*S1, *R1* and *R2* alleles; laboratory (n = 5), and field-caught mosquitoes (*6*) (n = 9) indicated by countries and accession numbers were included in the alignment. In total, the analyses involved 212 amino acid sequences in MEGA6 (p-distance parameters, 1000 replicates, 50% cut-off). Color codes of the tree branch represent alleles as follows: Green, *S1*; light green, *S2*; blue; *R1*; light blue, *R2*; and pink, *R3*.

The Neighbor Joining (NJ) phylogenetic analyses confirmed a clear separation of *TEP1* alleles into *S and *R clades. *TEP1*S* further splits to form two clusters of *S1* and *S2* subclasses. For *S1* alleles, no clear pattern of species-species or geographic clustering was evident. The majority of *S2* alleles originated from Cameroon samples, and showed apparent divergence from the only one *S2* from western Kenya and from the laboratory *G3* strain. *TEP1*R* segregated into three clusters of *R1, *R2* and *R3*, with *R3* at the root of the clade.

Next, *TEP1* sequences were separated into haplotypes using TCS 1.21 software (144). The software infers gene genealogies based on statistical parsimony, and depicts mutational steps (nucleotide substitutions) that separate or connect different groups of haplotypes. **Fig. 2-13**A, left panel, highlights a simplified DNA nucleotide genealogy network (not to scale) of major haplotypes across Africa. The NJ tree in **Fig. 2-13**B, displays relatedness of all the haplotypes shown in **Fig. 2-13**A.

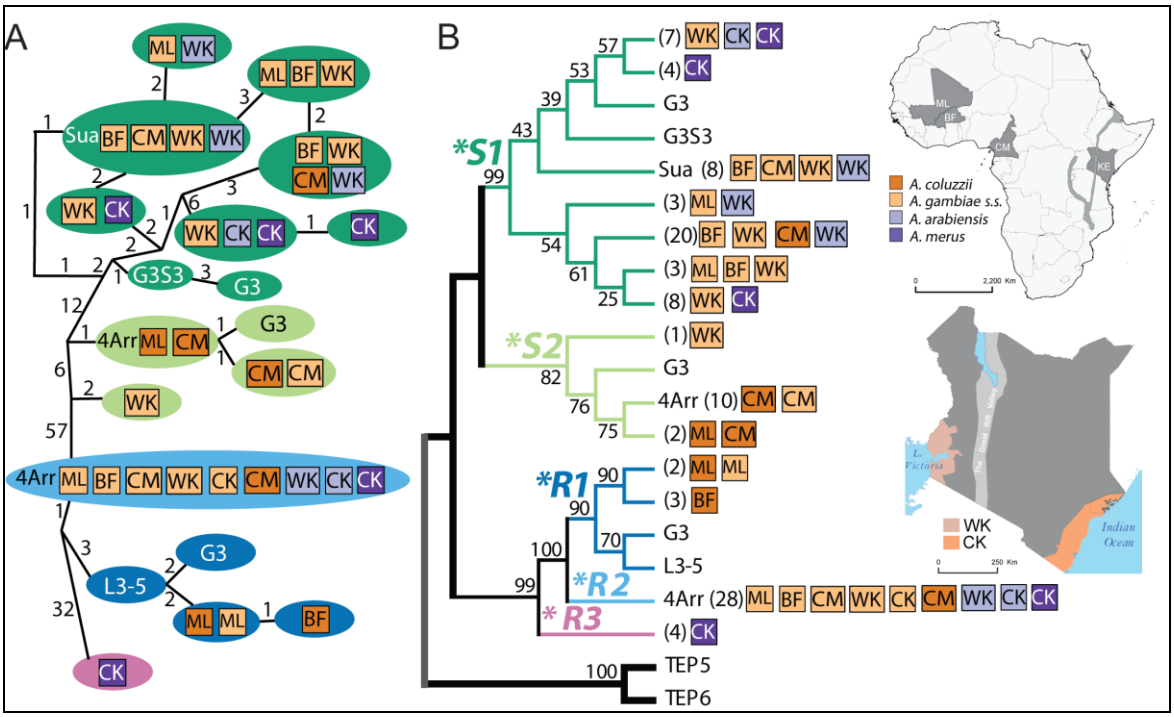


Fig. 2-13. Genealogy network and geodiversity of *TEP1* haplotypes.
(A) Schematic representation (not to scale) of genealogical network (left panel) of major haplotypes of the *TED* sequenced alleles that were inferred by the TCS software. Each oval shapes groups represent a haplotype composed of the species/strain and/or country of origin. The numbers in the left panel between each haplotype cluster represent number of nucleotide mutational steps or SNPs separating them.
(B) Neighbor joining tree (right panel) of the same haplotypes. The numbers in the right panel shown in brackets represent the actual number of individual alleles sharing the same haplotype. Abbreviations of the sampling sites or countries are as before. In both panels, different laboratory mosquito strains whose allele sequences were included are G3 (**R1, *R2* and **S2*), G3S3 (**S1*), Suakoko (**S1*), 4Arr (**R2* and **S2*), and L3-5 (**R1*).

The *TED* genealogy network and phylo-geography of *TEP1* haplotypes led us to suggest that: 1) *TEP1* alleles cluster independent of species and geography; 2) **R2* is the most conserved allele as revealed by the highest number (28) of shared haplotypes across the countries and species; 3) **R2* and **S1* are the most widespread haplotypes in all countries and in all species; 4) the major **R* and the **S* alleles groups are separated by the highest number of mutational steps (57), whereas the lowest number of mutational steps (>10) was observed within allelic subclass haplotypes; 5) based on the *TED,* the newly identified **R3* allele is closer to **R2* (33 steps) than is to **R1* (35 steps).

2.4.8 *TEP1*R3* allele displays unique amino acid substitutions

To assess the features of *TEP1*R3* allele, its sequences was translated *in silico* to amino acid residues. Amino acid substitutions within the *TED* were compared against the sequences from this study and those from other reports (*5*, *8*, *35*, *76*, *77*) (**Fig. 2-14**). Strikingly, **R3* substitutions were localized in the hypervariable loops i.e. loop before α-helix 4 (pre-α4 loop), catalytic loop and ß-hairpin loop (**Fig. 2-14**A).

Amino acid residues E1043 and D1058 in the ß-hairpin loop and K966, E970 and E974 in the catalytic loop were conserved in **R1, *R2* and **R3*. The **R1* and **R2* forms shared the same amino acid residues with **R3* except three **R1* to **R3* substitutions at N919G, N937K and V946M, and two **R2* to **R3* substitutions at T918S and A936S.

Indeed, *TEP1*R1, *R2* and **R3* also shared conserved residues at the pre-α4 loops L914, T917, T918 and G920. Moreover, *TEP1*R3* shared conserved residues with **S* forms at positions S918, N936, S940, S960, T991, A1005, and N1012. However, it featured private amino acid substitutions outside the catalytic loop. These are M1021 at pre-α8 and positions R1065 and K1067 at α-helix 10.

Relatedness of **R3* to both the **R* and **S* forms at the TED region was unexpected, therefore I extended the analyses to introns and exons of the full-length gene (**Fig. 2-14**B). For this, allele full-length genomic sequences were amplified and sequenced from **R3/R3* (n = 2), **R1/R1* from Mali (n = 1), **R2/R2* from Mali (n = 1), **R2/R2* from Kenya (n = 2) and **S1/S1* Kenya (n = 2).

In this analysis of homozygote mosquitoes, I used six primer pairs to amplify the full-length gene in overlapping PCR fragments (see Materials and Methods). While all fragments from **R3* from Kenya, **R2* from Mali and **R1* from Mali were successfully amplified and sequenced (Appendix **5**; Appendix **6**), I failed to amplify **R2* from Kenya using the pair of primers for the third fragment.

TEP1 gene has 11 intronic segments (Appendix **5**), which were trimmed off for analysis of the coding sequence. To visualize sequence alignments, NJ trees of the full-length nucleotide and amino acid sequences were constructed, similar tree topologies of allele segregation to those in **Fig. 2-12** and **Fig. 2-13**, were obtained (**Fig. 2-14**C).

Amino acid modifications outside the TED region were assessed by aligning the full-length coding sequences of the **R3* form together with those of **R1, *R2, *S1* and **S2* (Appendix **6**). Additional **R3* private amino acid modifications were found in MG3, MG5, MG7 and MG8 domains (**Table 2-10**; Appendix **5**; Appendix **6**).

Table 2-10. *TEP1*R3* full-length amino acid modification.

		Amino acid substitution	
Position	Domain	**R/*S*	**R3*
266	MG3	D	E
289	MG3	D	N
489	MG5	R/K	M
742	MG7	Y	C
759	MG7	L	F
760	MG7	I	V
1005	TED	E/N	A
1021	TED	I	M
1065	TED	G/N	R
1067	TED	T	K
1223	MG8	Q/E	K

Next, polymorphism in the introns was evaluated by joining all the trimmed 11 intronic sequences to form a continuous contig of 807 bp sequence, which was used for the NJ alignment. Sliding window analyzes showed a marked variation in nucleotide diversity along the gene sequence. The diversity was low in introns 1 to 7 (0-0.05) and higher in the introns 8 to 11 (0.05-0.23), suggesting that high levels of genetic diversity around *TED* in the coding regions are in concert with high levels of genetic diversity in the noncoding sequences (**Fig. 2-14**B). In addition, the **R3* allele has nucleotide substitutions that decorate the introns 1-6 and 8-9, substitutions and insertions in the intron 7, only insertions in intron 11, but intron 10, which was allele-specific, did not have any modifications compared to **R1, *R2, *S1* and **S2* alleles (**Fig. 2-14**C).

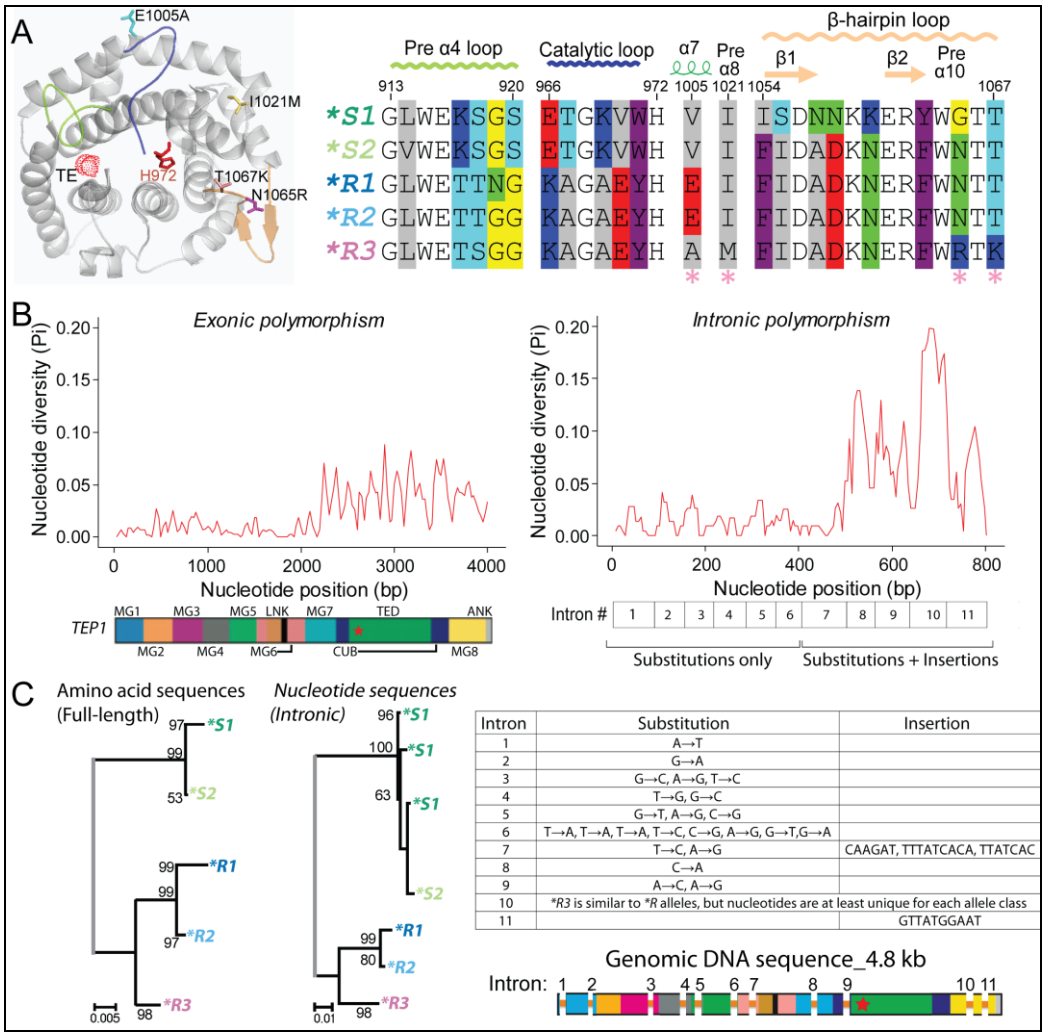


Fig. 2-14. Overview of unique *TEP1*R3* amino acid and nucleotide sequence variability.
(A) Comparison of *TEP1*R3* amino acid residues with those of other alleles at the TED. *TEP1*R3* private SNPs are mapped on TEP1*R1 structure. The conserved catalytic histidine (H972) and *TEP1*R3*-specific substitutions are shown as colored sticks. The TED peptide is shown as a cartoon in grey and TE (thioester) as a red bubble. Hypervariable pre-α4, catalytic and β-hairpin loops are shown in green, blue and light yellow colors respectively. Amino acid alignment displays comparison of *TEP1*R3* residues with that of other alleles in pre-α4 (913-920), catalytic (966-972) and ß-hairpin (1054-1067) loops. The pink asterisks below the alignment indicate four positions of *TEP1*R3* amino acid substitutions (A1005 and M1021 found in most **R3* haplotypes, and R1065 and K1067 found in all **R3* haplotypes). Amino acid residue groups are colored as follows; AVLIM (gray), G (yellow), FY (purple), N (green), ST (Turquoise), DE (red), KR (Blue). The TEP1*R1 structure was adapted from Le *et al.,* 2012 (*76*).
(B) Sliding window plots compare nucleotide polymorphism between the sequences of exons (left panel) and introns (right panel). Below the plots are the corresponding TEP1 domain map and 11 introns respectively. Higher intronic nucleotide diversity occurs in the last four introns found in regions flanking the *TED*. Higher nucleotide diversity in exons at these parts within and in neighborhood of the *TED* that correspond to the C-terminus of the protein. Sequences include two **R3* full-length sequences. See Appendix **5** and Appendix **6**.
(C) Visualization of the alignment showing *TEP1* allele separation between coding and intronic sequences. A neighbor-joining tree (first tree) using amino acid full-length sequences and nucleotide maximum likelihood (second tree) based on only intronic sequences. The NCBI Genbank accession numbers of *TEP1* full-length sequences are MF098568 to MF098592. The table highlights **R3* modifications in the introns. Below it, is a schematic map depicting the positions of the 11 introns in the genomic *TEP1* sequence. For details, see Appendix **5** for complete sequence alignment.

2.5 Discussion

Evolution of immune genes is functionally constrained by predominantly purifying selection in order to conserve their role in immunity, hence detecting evidence of sites under positive selection is very rare (*100*). Fixed *TEP1* allele-specific SNPs at the *TED* region provide clear inter-allelic differences (*5*, *76*, *77*). The *TED* allele-specific SNPs that were used in PCR-RFLP to genotype *TEP1* in mosquito populations were robust in resolving all the *TEP1* alleles (*80*). Importantly, the method may become a useful tool for high throughput *TEP1* genotyping that can be easily performed in the field.

All the four *TEP1* alleles (**R1, *R2, *S1* and **S2*) previously identified in laboratory strains are also found in the natural *A. gambiae s.l.* complex malaria vector in Africa (*5*, *8*, *35*). Taking into account the newly identified *A. merus* **R3* allele, a full panmixia scenario predicts 15 possible *TEP1* genotypes. However, our data identified a restricted number of genotypes across Africa. This is partially due to geographical separation between the species that carry geographically restricted alleles (for example, **R1* in Mali, **R3* in *A. merus* in coastal Kenya, and **S2* in Cameroon). Therefore, we identified only 11 (5 homozygous and 6 heterozygous) *TEP1* genotypes across Africa.

The observed variation in the biogeographic distribution of *TEP1* alleles and genotypes allowed categorization of genotypes into generalists and specialists. Thus, **S1/S1* and **R2/S1* as generalist genotypes since they were identified in all the countries and in all species, whereas **R1/R1, *R3/R3, *R3/S1, *S1/S2* and **S2/S2* were categorized as specialist (or restricted) genotypes as they show either geographical- or species-specific restriction. Note that **R1/S1, *R1/S2* and **R2/S2* were found at very low frequencies (<0.05) in regions where specialist alleles exist. Some *TEP1* genotypes, such as **R1/R2, *R1/R3, *R2/R3* and **R3/S2* were absent in the mosquito populations, and therefore were categorized as undetected genotypes. Surprisingly, **R* alleles were only found in homozygosity or in heterozygosity with **S*, suggesting potential functional incompatibility between **R* alleles.

In line with the previous reports (*8*, *35*), the **R1/R1* genotype was only found in Mali and Burkina Faso. This could result from genetic restriction and adaptation to arid habitats in the Sahel region, or some drier Savanna ecotypes. The **S2/S2* genotype which was recently associated with higher male reproductive fitness (*80*), was found only in low frequencies in Mali and Kenya. Enrichment of **S2* allele in Cameroon indicates that some selection factors may be at work, however, the nature of these

factors remains to be investigated. Interestingly, the generalist **S1/S1* and specialist **S2/S2* mosquitoes present in Cameroon both contribute to the maintenance of the **S1/S2* genotype, as suggested by a high number of Cameroon-restricted **S1/S2* heterozygotes.

Although, the frequency (~0.1) of **R3/R3* and **R3/S1* genotypes was very low in Kenya, these mosquitoes were sampled inside residential homes using manual aspiration method. It is important to highlight here that the reason for the low mosquito numbers in Kenya was due to the nature of the sampling plan designed to catch blood fed adult mosquitoes in human dwellings. Getting at least one mosquito in a given household was precious!

This study uncovered a well-refined architecture of geographic- and species-*TEP1* genotype variation. The departure of *TEP1* genotypes from the HWE expectation suggests that one or more of the HWE assumptions are violated, and indicates action of some selection forces that drive ecological adaptations in *A. gambiae s.l.*. High values of F_{ST}, global fixation indices are indicative of genetic differentiation in all populations, suggesting presence of defined local genetic structures at the *TEP1* locus. Moreover, high inbreeding coefficients in some mosquito populations are indicative of extensive inbreeding between the mosquitoes. The **R1/R1* genotypes are present at near-fixation in *A. coluzzii* mosquitoes in Mali and Burkina Faso but absent from Cameroon (8, 35). This observation may suggest that as a result of adaptation to different ecological niches, *A. coluzzii* has undergone significant diversification due to geographical and reproductive isolation (8, 35). Another important factor may be inter-species competition in the breeding sites, for example competitive exclusion of *A. coluzzii* has been observed in wet areas in the absence of *A. arabiensis* and *A. gambiae s.s.* (46).

The presence of rare **R1/S1* genotypes in *A. coluzzii* and in *A. gambiae s.s.* and of **R1/S2* in *A. coluzzii* populations may suggest a moderate reciprocal gene flow between the species. Alternatively, **R1* might have been selectively eliminated from *A. gambiae s.s.* and the few examples represent some vestigial evidence. Importantly, the absence of **R1/R1* genotypes in *A. gambiae s.s.* populations may be an indication of fitness cost associated with this **R1* allele in *A. gambiae s.s.* By extension, the occurrence of **R1/R1* genotypes in natural hybrids between the *A. coluzzii* and *A. gambiae s.s.* species, suggests that genetic background of *A. coluzzii* species benefits **R1/R1* genotype.

Most *A. merus* species were found in salty-water environments in Malindi in coastal Kenya, where no *A. gambiae s.s.* was found likely due to competitive exclusion. Therefore, diverse ecological settings provide core niches for diversification and adaptation of anopheline species (43). In the salty-water biotopes, the generalist **S1/S1* genotype is predominant, suggesting that it is not subject to demographic and ecological effects in most species, and is under strong purifying selection driving its conservation.

Cameroon-restricted **S1/S2* genotypes were found both in *A. coluzzii* and *A. gambiae s.s.*, implying that the environmental conditions prevailing in Cameroon drive selection for **S1/S2* genotypes. Yet, another consistent observation in Cameroon was the higher frequency of **S1/S2* genotypes in *A. coluzzii* relative to **S1/S2* genotypes in *A. gambiae s.s.,* which may imply that functional species-specific differential selection for **S1/S2* genotype occurs. This raised the question whether **R2/S1* genotype in *A. gambiae s.s.* could be more competitive than **S1/S2,* since the relative frequency **R2/S1* genotype was higher in *A. gambiae s.s.* than in *A. coluzzii* populations. This apparent negative correlation between **R2/S1* and **S1/S2* genotypes and between the two species may be interpreted as a signature of competition between the two genotypes at the habitat level. Moreover, it appears that this trade-off between **R2/S1* and **S1/S2* genotypes is species-specific. In contrast, the specialist **S2/S2* genotypes were identified in all species at low frequencies in Cameroon, Mali and Kenya. This suggests that **S2/S2* is more widespread than **S1/S2* genotypes.

Mainly, the generalist **R2/R2, *R2/S1* and **S1/S1* genotypes maintain **R2* and **S1* alleles observed across all countries and in all species. Since all *A. gambiae s.l.* species feature **S1/S1* and **R2/S1* across African ecological ranges, this may be indicative of some advantages offered by **R2* and **S1* alleles. Importantly, this allelic combination may point to conserved but yet unknown biological functions intrinsic to the **R2/R2, *R2/S1* and **S1/S1* genotypes, and hence the **R2* and **S1* alleles.

The generalist **R2/R2,* specialist **S2/S2* and **S1/S2,* and rare **R2/S2* genotypes maintain **R2* and **S2* alleles in most species. Strikingly, the unique switched pattern in genotype frequency in **S1/S2* and **R2/S1* genotypes between *A. coluzzii* and *A. gambiae s.s.* in Cameroon may suggest that **R2* and **S2* alleles compete with each other to confer competitive fitness and/or advantage to the species. It is inferable from the data that the generalist **S1* allele is compatible with all the other alleles as exemplified by the presence **S1* carriers i.e. **R1/S1, *R2/S1, *R3/S1* and **S1/S2*

genotypes, hence underpins unknown functional fitness advantage underlying this compatibility.

Studies have suggested that about 5000 years ago, gene flow and recent expansion in malaria vectors may be due to active migration including passive transportation through human activity (*32*, *119*, *153*). Yawson *et al.* (2007) (*99*), observed absence of hybrids in Ghanaian *A. coluzzii* and *A. gambiae s.s.* populations and no species-specific genetic differentiation but implicated strong genetic differentiation among diverse ecological areas. In far-West Africa, especially in Guinea Bissau, extensive hybridization between the two species suggests that this region is a central hybridization zone (*8*, *39*, *115*, *130*). Our study found evidence of low gene flow rate, as exemplified by the paucity of observed hybrids between *A. coluzzii* and *A. gambiae s.s.*. Due to low frequency of hybrids, it will be challenging to determine the natural mating potential of these hybrids and their competence for malaria transmission.

Most generally, striking differences were observed between *A. coluzzii* genotypes in Mali, Burkina Faso and Cameroon, strongly suggesting environmental factors that drive the selection of these genotypes. Previous studies suggested that differences in ecological niches of the malarial vector karyotypes are particularly important drivers of genetic differentiation (*102*). Variation in ecological niches pushes species to adapt by developing special discrete phenotypes as seen in genetic structuring of *A. arabiensis* based on mtDNA *ND* locus (*154*). In another study, it was observed that local *A. arabiensis* populating in more arid or hottest ecotypes have darker or melanic skin color and bigger body size as compared to *A. arabiensis* from more humid or cooler regions. Moreover, these melanic forms are more frequent and persistent throughout seasons of dry spell (*154*, *155*). However, their study did not interrogate the *TEP1* genotypes of these species.

Taken together, our data suggest that the distribution of generalist (wide spread) and specialist (restricted) genotypes may be associated with ecological adaptation to diverse selective constraints such as hosts, pathogens, and climatic stress. (*8*, *35*). The **R1/R1* genotype in *A. coluzzii* in hybridization zones in far-West is characterized by extensive outbreeding with other *A. gambiae s.s.* genotypes, thus, demonstrates a correlation between *TEP1* locus and eco-geographical partitioning of *A. coluzzii*. Previous studies using chromosomal inversions or microsatellite markers had reported separation of *A. coluzzii* into subpopulations. For instance, arrangements on the 2R

chromosome had revealed three *A. gambiae* reproductive units, called Bamako, Mopti and Savanna, where complete reproductive isolation was observed between Bamako and Mopti but incomplete isolation between Bamako/Mopti and Savanna (46). Interestingly, *TEP1* localization on chromosome 3L suggests that it is not linked to the inversions that separate Bamako, Mopti and Savanna forms. The correspondence between *TEP1* polymorphism and these forms remains to be elucidated.

Deviations from the HWE expectations suggest existence of environmental forces (such as climatic or local factors) that operate at the *TEP1* locus. As *TEP1*S1* and **R2* are the most widespread alleles present today across the African countries, they likely are representing the most conserved ancestral forms spanning over many million mosquito generations. The male fertility-associated *TEP1*S2* alleles (80) are mostly selected in Cameroon but are kept at very low frequencies in Mali and Kenya.

While the **R2* and the **S1,* and to a lesser extent the **S2* alleles are conserved across Africa, the *A. coluzzii *R1* and the *A. merus *R3* private alleles are not only the most diverse, but also show geographical- and species-specific pattern. It can be inferred that strong directional selection is a driver of this geographical- and species-specific selection pressures, hence this pattern of local adaptation.

Moreover, *TEP1*R1* and **R3* and, to some extent, **S2* alleles suggest divergent evolutionary paths that *A. gambiae s.l.* developed to confront certain biotic and/or abiotic ecological constraints specific to their ecological niches. These differences in allele and genotype selections in nature may be, in part, due to differences in the choice and stability of breeding habitats (118) or prevailing ecological climates of which local factors and geographical isolation significantly render differences in the breeding ecology, hence speciation, diversification and selection of fitter genotypes (98, 118).

The genealogy network and phylo-geographic analyses confirm that *TEP1* alleles have been maintained in the populations of malaria vector throughout Africa (6, 8, 35), with substantially shared polymorphism among species. Although there are more haplotypes within the *TEP1*S1* alleles (partly due to the oversampling of the sequenced **S1* alleles), **S1* haplotypes cluster discretely from the haplotypes of other alleles. This may be a result of differences in accumulation of diverse ecotype-specific mutations that do not switch allele classes, hence the conservation in all countries in all mosquito species. The majority of **R2* sequences share a single haplotype, highlighting the highly-conserved nature of this allele. According to the coalescence theory (156), the

striking conserved nature of **R2* and, to a lesser degree of **S1* haplotypes, and the propensity of both to form heterozygotes, further suggests that they represent the most maintained ancestral alleles whose polymorphism in African *A. gambiae s.l.* populations is functionally constrained.

The striking similarity of TEP1*R3 amino substitutions with those of both TEP1*R and TEP1*S forms, led us to suggest that the new allele may be or is closely related to an ancestral form of both **R2* and **S* alleles. Analyses of the full-length **R3* genomic sequences revealed the extent of conservation of this new allele. First, uneven nucleotide substitutions and/or indels were identified in all the introns. Uniquely, the intron 10 appeared to be allele-specific and did not have any nucleotide modifications compared with the **R1, *R2, *S1* and **S2* alleles. Secondly, high nucleotide diversity was detected in the introns 8 and before *TED* and introns 10 and 11 in MG8 suggesting functional contrains in both the introns and the exons. Thirdly, seven additional private amino acid substitutions in other domains may indicate that they evolve together and likely to be important for TEP1*R3 function, which is currently unknown.

2.6 Conclusion

The marked divergence between *A. coluzzii* in Mali and Burkina Faso from *A. coluzzii* in Cameroon suggests partitioning of this species into ecotypes directly or indirectly linked to *TEP1* locus. Therefore, structure of mosquito populations at the *TEP1* locus offers an important tool in assessing gene flow radiation and genetic dispersal of malaria vectors. This study has successfully used evolutionary genetics and molecular tools to understand *TEP1* genetic diversity and population genetic structure of malaria vector populations in field ecological surveys in West Africa, Central Africa, and East Africa. For the first time, this study mapped *TEP1* alleles and genotypes in four malaria vector species from the four African countries. The SNPs used for *TEP1* genotyping were robust and may be harnessed as genetic or molecular markers for high throughput *TEP1* genotyping. The data discussed here demonstrate that the genetic variation in *TEP1* locus shape the population genetic structure in local malaria populations. The *TEP1* alleles and genotypes are structured according to geography and local vector species. The *TEP1* genetic diversity that was observed matches different African climatic zones suggesting that it enables the mosquitoes to adapt to the prevailing environmental conditions.

Chapter 3

Impact of *TEP1* variation on development of *P. falciparum*

Impact of *TEP1* variation on development of *P. falciparum*

3.1 Summary

Little is known about the impact of *TEP1* allelic variation on *P. falciparum*. The aim of the project described in this chapter was to examine the impact of *TEP1* variation on *P. falciparum* development. To this end, I established a mosquito line that contained three *TEP1* alleles: **R1, **S1* and **S2*. I compared the outcomes of infections with human *P. falciparum* and murine *P. berghei* between *TEP1* genotypes. Interestingly, *TEP1*S1/S1* and *TEP1*S2/S2* appear to be equally susceptible to *Plasmodium* infections. Rearing of the *TEP1*R1/R1* homozygous mosquitoes was challenging. Due to the high mortality rates, only few mosquitoes homozygous for **R1/R1* were used, preventing meaningful statistical analyses. Nevertheless, these preliminary results indicate a trend in resistance of *TEP1*R1/R1* to *P. falciparum* infections as compared to *TEP1*S* mosquitoes. High mortality of **R1/R1* mosquitoes suggests that *TEP1*R1* may be conditional lethal allele, and therefore, future functional experiments with the **R1/R1* mosquitoes should be done in their natural ecology in Africa.

3.2 Introduction

The highly polymorphic nature of thioester-containing protein 1 (TEP1) in the *A. gambiae* mosquito mediates a powerful immune response against *Plasmodium* parasites during midgut invasion (4, 135, 157). Mosquitoes bearing different *TEP1* genotypes have distinct phenotype variation in the mode of clearance of *P. berghei* ookinetes (5). In *TEP1*R1* mosquitoes, the ookinetes are bound with faster kinetics and dead parasites are cleared by both lysis and melanization, compared to *TEP1*S* alleles which mediate killing by lysis only (135). *TEP1*R1* mosquitoes display relatively lower *P. falciparum* (*3D7* strain) infection rates than *TEP1*S* (35). However, in *TEP1*R1* homozygous strains, the Brazilian *7G8* and the African *NF54* (the parental strain of *3D7*) *P. falciparum* differ greatly in the parasite killing and clearance, suggesting that *P. falciparum* infectivity is dependent on the genetic background of the parasites (158). Notably, *P. falciparum* parasites express *P. falciparum* surface protein (Pfs47) on the ookinetes to enable them to efficiently evade the immune system of the mosquito and mature into the oocysts in the midgut (159). Recently, Pfs47 protein was shown to mediate the survival of the ookinetes by disrupting the mosquito JNK signaling pathways against the parasite (160, 161). Conversely, efficiency of Pfs47 in mediating

immune evasion varies between African parasites and other parasites suggesting that genetic adaptation of the parasites, genetic differences between the mosquitoes and parasite-vector compatibility may play a role (*159*, *162*). Sequencing of *P. falciparum* isolates in Africa and of culture lines failed to detect a correlation between susceptibility to TEP1-mediated killing and *Pfs47* genotypes (*104*). These observations suggest that TEP1 is an important anti-plasmodial factor for *P. falciparum* transmission, and that genetic variation at *Pfs47* locus is not the sole mediator of *P. falciparum* evasion of TEP1 immune responses (*104*). However, their study did not address the impact of *TEP1* polymorphism on *P. falciparum* development.

In this project, I examined whether *TEP1* polymorphism impacts *Plasmodium* development. To this end, parasite loads and prevalence of *P. berghei* and NF54 *P. falciparum* infections in laboratory mosquitoes bearing *TEP1*R1, *S1* and **S2* alleles were determined. First, mosquito line, herein named *H3T1*, bearing *TEP1*R1, *S1* and **S2* alleles was established. Next, the line was infected with murine *P. berghei* and human *P. falciparum* parasites to assess the impact of *TEP1* genotypes on the infections. Results show that *TEP1*S1/S1* and *TEP1*S2/S2* mosquitoes appear to be equally susceptible to *Plasmodium* infections. Interestingly, a trend for *TEP1*R1/R1* genotype showing resistance to *P. falciparum* infections was observed. However, the *TEP1*R1/R1* mosquitoes suffered high mortality rates from one generation to the next, therefore, only low numbers of **R1/R1* homozygotes were tested in the infection experiments.

3.3 Materials and Methods

All the materials (consumables and biological), equipment and software that were used in this chapter are listed in Appendix **1**.

3.3.1 *Plasmodium berghei* strain and maintenance

P. berghei ANKA (PbGFPcon) (*163*), murine parasite strain that constitutively expresses a green fluorescent protein (GFP) was maintained *in vivo* using CD1 mice. Parasitaemia of the infected mouse was assessed by FACS.

3.3.2 *Plasmodium falciparum* strains and maintenance

All the Standard Operating Procedures of culturing *P. falciparum* strains were followed according to the standard settings of security level 2 (S2) laboratories as per the German national regulations of the Landesamt for Gesundheit und Soziales

(LAGESO). The standard *P. falciparum* strain NF54 (164) and strains; *HBC (165), NF54HT-GFP-luc (166) , NF165 (104)* and *7G8* (158, 167) were routinely cultured.

3.3.3 Mosquito strains and maintenance

A. gambiae mosquitoes*, Mut6* (*TEP1*S1/S1* mutant) (138), *Ngousso* (*TEP1*S1/S1*) (168), *7b* (*TEP1*S1/S1* knockdown line) *(78)*, 4Arr (MR4) (5), L3-5 (*TEP1*R1/R1,* refractory strain to *P. berghei*) (167, 169) and *G3* (https://www.beiresources.org/Catalog/livingMosquitoes/MRA-112.aspx) were used in this study. The G3, is herein referred to as *"H3T1"* or '*H3T1 WT',* had three *TEP1* alleles: **R1*, **S2* and **S1*. Both *7b* and *Mut6* were used as TEP1-deficient control strains. *Mut6* strain is *TEP1*S1/S1* homozygous for a premature stop mutation at the alpha helix region connecting the CUB and TED that rendered the mosquito hyper-susceptible to infections by *P. berghei* (138). Mutations in *TEP1* were generated by TALENS technology resulting in the deletion of 3 endogenous amino acids and the insertion of 6 exogenous amino acids (138). Importantly, *Mut6* was used in intercrosses with the *H3T1* strain to introduce a mutant allele in order to develop *TEP1*S1* hemizygous mosquitoes.

Adult mosquitoes were fed daily on 10% sucrose solution and maintained under standard insectary conditions at 28 ± 2 °C and 75-80% relative humidity in a 12 h day/night cycle. For oviposition, female mosquitoes were blood fed for 15 min on human type O+ blood (HAEMA, Berlin) using a standard membrane feeder (SMF). After 48 h, an egg dish (a damp funnel-shaped filter paper placed over a jar filled with sterile water) was placed in the cage for egg-laying. Eggs were collected 24 h later and floated in plastic pans containing sterile water supplemented with 0.1% NaCl. Upon hatching, larvae were fed on ground Tetrafin Flake Fish food (Tetra, Germany). Pupae were collected in water-containing jars and transferred into standard cages (30×30×30 cm) (http://bugdorm.megaview.com.tw/bugdorm-43030f-insect-rearing-cage-32-5x32-5x32-5-cm-pack-of-one-p-241.html) and allowed for adults to emerge.

3.3.4 Breeding of the *H3T1* mosquito strain and balancing *TEP1* allelic composition and genotype frequencies

The original *H3T1* colony was genotyped for *TEP1* to assess the *TEP1* genotype frequencies. Of the three alleles, *TEP1*R1* had the lowest allele frequency. Therefore, a selection procedure was set in place to establish a *H3T1* colony with equal *TEP1* allelic frequencies. For this, *TEP1*R1* males (n = 28) were selected and allowed to freely mate

with 7-day-old virgin females (n = 165). The F1 offspring were maintained to adulthood and self-crossed to produce F2 progeny. The F2 offspring were genotyped for *TEP1,* and then male and female mosquitoes with desired *TEP1* genotypes were allowed to mate and give rise to F3 progeny. The subsequent generations were assessed for *TEP1* genotype and allelic frequencies. The colony was enriched accordingly for the *TEP1*R1* alleles whenever a declining shift was observed in its frequencies (*35*). However, the original **R1* allele was not successfully stabilized, thus an alternative **R1* allele from L3-5 refractory mosquitoes was introgressed into the *H3T1* strain. I also tried to introgress *TEP1*R2* allele from the 4Arr strain obtained from MR4 into the *H3T1* colony. However, both the introgression experiments were unsuccessful.

3.3.5 *MH3T1* reciprocal crosses

To carry out reciprocal crosses between the *mut6* and *H3T1* mosquito lines, pupae from each line were sexed under a microscope by the examination of terminalia. Males of *mut6* and females of *H3T1* were merged as pupae into a mosquito cage and allowed to emerge to adulthood. Similarly, males of *H3T1* and females of *mut6* pupae were merged into another mosquito cage and allowed to emerge to adulthood. Female adults were blood-fed between 5- and 7-d-post-emergence. The mosquitoes were provided with egg dishes for oviposition on 48 h post blood-meal acquisition. The resulting egg dishes were merged into one floating pan and allowed to hatch to larvae. Mosquitoes were reared to adulthood and used for *P. berghei* infection experiments.

3.3.6 DNA extraction

Mosquito genomic DNA for PCR analyses was extracted from mosquito legs by lysis in 40 µl of squashing buffer (200 µg/ml proteinase K, 10 mM Tris-Cl pH 8.2, 1 mM EDTA and 25 mM NaCl) and followed by 1 h incubation with proteinase K at 37 °C with subsequent 5 min inactivation at 95 °C. The lysates were cleared by high-speed centrifugation to pellet the proteins and tissues that may inhibit the PCR. Alternatively, a leg was used directly as the DNA template in the PCR reaction.

3.3.7 *TEP1* genotyping

Nested-PCR RFLP and PCR-based genotyping strategies (Chapter 2) were used to genotype *TEP1*. The reciprocal crosses between the *mut6* and *H3T1* strains were genotyped using the 1034±1 bp PCR-RFLP genotyping method. See **Table 3-1** for the expected sizes of RFLP fragments.

Table 3-1. RFLP fragment (bp) expected from genotyping the *Mut6-H3T1* genetic crosses.

	Restriction enzyme and expected fragment sizes	
TEP1 genotype	*Bam*HI+*Nco*I	*Bse*NI
S1/S1	146, 887	811, 222
S1/S2	146, 887	1033, 811, 222
S2/S2	146, 887	1033
R1/R1	674, 360	1034
R1/S1	674, 360, 146, 887	1034, 811, 222
R1/S2	674, 360, 146, 887	1034*, 1033*
S1m/S1m	1043$	821$, 222
R1/S1m	674, 360, 1043$	1034, 821$, 222
S1/S1m	146, 887, 1043$	811*, 821*$, 222
S2/S1m	146, 887, 1043$	1033, 821$, 222

*Fragments have close size ranges and would appear as one overlapped fragment.
$ The expected size of *S1m* allele is 9 bp more due to the targeted deletion of 8 endogenous bases and subsequent insertion of 17 exogenous bases at the *Nco*I site (*138*).

3.3.8 Experimental infections

3.3.8.1 *P. berghei* infections

Female mosquitoes (3- to 5-d-old) were infected with the GFP-expressing rodent malaria parasite (*P. berghei* ANKA, PbGFPcon) (*163*). The mosquitoes were starved for 6 h prior to a 20 min blood feeding on an anaesthetized *P. berghei*-infected CD1 mouse by injection of the mice with a mix (120 µl) of xylazine and ketamine in line with the national regulations by the LAGESO. Fully blood-fed mosquitoes were selected and maintained by feeding daily on 10% sucrose solution in S2 incubator set at 20 °C and 75-80% relative humidity in a 12 h day/night cycle for 10 d post-infection. Midguts were dissected in sterile PBS under a stereomicroscope. For oocysts counting, midguts were fixed in 4% paraformaldehyde for 30 min and washed 3 times in PBS. Midguts were mounted on a slide, stained with DAPI (Vector Laboratories, USA). The images were documented under an Inverted Fluorescence Microscope (Zeiss Axio Observer Z1).

3.3.8.2. *P. falciparum* infections

All infectious work on the human malaria *P. falciparum* parasite cultures and infection of mosquitoes was performed under controlled security level 3 (S3) laboratory conditions according to the standard regulations by the LAGESO, project number 411/08. In order to infect mosquitoes with the *P. falciparum*, the SMF system was used. Here, ~ 1-5% stage V gametocyte cultures of routinely maintained *P. falciparum* NF54 incubated with human blood cells (HAEMA, Berlin) were fed to the 3 to 5-d-old

females using the SMF stystem for 20 min at 37 °C. Unfed and partially engorged mosquitoes were removed from the cage and killed immediately using a vacuum aspirator into 70% ethanol. Fully engorged mosquitoes were maintained in the S3 conditions for 10 d at 26 °C and 75-80% relative humidity in a 12 h day/night cycle.

After 10 d post infection, mosquitoes were killed in 70% ethanol and washed 3 times in 1× PBS. Mosquitoes were dissected in sterile PBS under a stereomicroscope to remove the midguts for assessment of *P. falciparum* oocyst loads. To visualize and count the developed oocysts, the midguts were stained with 1% mercurochrome for 10 min, and observed under a bright field upright microscope (Leica DM2000 LED) equipped with a CCD color camera.

3.3.9 Analyses

R version 3.1.3 (2015) customized scripts and R-packages as in Chapter 2 were used to analyze the genotype and allele frequencies from the *TEP1* genotyping data (Appendix **2**). All statistical analyses were carried out in R statistical package version 3.1.3 (2015) (*141*). These include the assessment of normality of the data using Q-Q plots and Shapiro tests, analyses of oocyst numbers using the non-parametric Kruskal-Wallis test with pairwise Wilcox test and Bonferroni correction. Analyses of means of *Plasmodium* prevalence by one-way ANOVA on log-transformed data followed by the Tukey's HSD test. A sample of an R script is given in Appendix **7**. Figures were edited in Adobe Illustrator C5 and Adobe Photoshop C5.

3.4 Results

3.4.1 Establishment of the mosquito colony with balanced *TEP1* allelic and genotype frequencies

My aim was to examine the impact of *TEP1* alleles on *P. falciparum* development. However, the attempts to introgress **R2* allele from 4Arr strain (obtained from MR4) were unsuccessful.

Nonetheless, the availability of the *H3T1 A. gambiae* mosquito colony bearing *TEP1*R1*, **S1* and **S2* alleles was used for *Plasmodium* infection experiments. First, original *H3T1* strain was genotyped for *TEP1* to assess allelic frequencies and genotypes (**Fig. 3-1**, experiment 1). Strikingly, I did not detect any homozygous *TEP1*R1/R1* (n = 96), although **R1* allele was maintained by the heterozygote **R1/S1* and **R1/S2* genotypes (2% each) (**Fig. 3-1**A). **R1* showed the lowest frequency (2%),

while frequencies of **S2* (41%) and **S1* (57%) were similar (**Fig. 3-1**B). Therefore, it was important to increase the frequency of **R1* allele in *H3T1* before performing *Plasmodium* infections.

All mosquitoes carrying **R1* alleles were self-crossed to generate F1 progeny with enriched **R1* alleles. The F1 progeny were reared to adulthood and self-crossed to produce F2 generation (**Fig. 3-1**, experiment 2). The F2 progeny was genotyped and **R1* individuals were used for crosses to further enrich the frequency of **R1* and of **R1/R1* genotypes. The colony was maintained with regular genotyping to assess frequencies of genotypes and alelles and accordingly enrich it with **R1* alleles (**Fig. 3-1**, experiments 3-7).

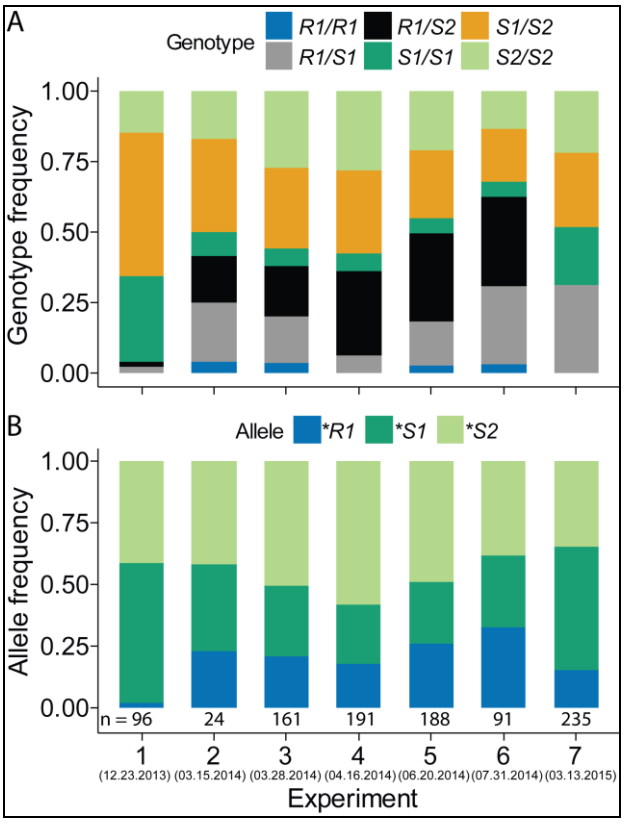


Fig. 3-1. Equilibration of *TEP1* allelic and genotype frequencies in the *H3T1* strain.
(A) Overview of *TEP1* genotype frequencies of the *H3T1* upon with *TEP1*R1* enrichment in the colony. The x-axes represent the number of experiment. Experiment 1 depicts genotype and allelic frequencies in the original colony.
(B) *TEP1* allelic frequencies with constant enrichment of *TEP1*R1* allele in the *H3T1* colony. Numbers below indicate the number of the genotyped individuals in each experiment. Below each experiment number is a specific date (enclosed in brackets) when the genotyping of each colony was done.

All my attempts to establish the *H3T1* colony with equilibrated genotype frequencies were unsuccessful (**Fig. 3-1**). The homozygous **R1/R1* mosquitoes always displayed the lowest frequencies, resisting my selection strategy. Interestingly, self-crossing of **R1* individuals increased allelic frequencies of **R1*, but not the proportion of **R1/R1* individuals, suggesting some degree of homozygote lethality of **R1* allele.

To counteract this challenge, a strategy of a continuous boosting of the colony with **R1* mosquitoes was adopted. However, this strategy required an extremely high

number of total mosquitoes (>600) to analyze per experiment. This high number of mosquitoes was not achievable and sustainable hence required a continuous optimization during the breeding cycles.

To avoid potential competition between **R1/R1* and other genotypes, I established homozygous lines of **R1/R1, *S1/S1* and **S2/S2* mosquitoes. Note that establishment of homozygous **R1/R1* took longer time than **S1/S1* or **S2/S2* homozygote colonies. Moreover, the **R1* homozygous colony experienced frequent crushing due to poor adult blood feeding and low numbers of laid eggs.

I tested another option to enrich the frequencies of **R1* alleles by introgressing **R1* allele of refractory strain L3-5 into the *H3T1* colony bearing **S1* and **S2* alleles. This strategy needed a considerably high number of generations (>6) involving the intercrosses to synchronize and equilibrate genetic background and to avoid the confounding impacts of founder effect. However regardless of the **R1* origin, **R1/R1* homozygotes were challenging to breed. Therefore, the *H3T1* colony with very low frequency of **R1/R1* genotypes (around the frequencies of **Fig. 3-1**, experiment 5 and 6) was used to perform *P. berghei* and *P. falciparum* infection experiments.

3.4.2 *MH3T1* mosquito colony establishment

To obtain hemizygote mosquitoes for assessing the impact of single copies of *TEP1*R1* or **S1* or **S2* on *Plasmodium* development, I set up reciprocal crosses between *H3T1* and *mut6* mosquito cohorts. The mutant *TEP1* allele is referred to as **S1m*. In principle, the F1 progeny should inherit one copy of *TEP1* allele from the *H3T1 WT* (wild type) and one from the *Mut6* (mutant). This allelic combination offered an opportunity to directly compare the impact of single *TEP1* alleles on *Plasmodium* development.

3.4.3 *P. berghei* infections of the *MH3T1* mosquito reciprocal crosses

To dissect the contribution of a single *TEP1* allele to resistance to *Plasmodium* infection (and to resolve the synergistic effects of two alleles in the mosquito), F1 generations of the *MH3T1* mosquitoes were infected with the *PbGFPcon* parasites (**Fig. 3-2**). Briefly, mosquitoes were allowed to feed on anaesthetized *PbANKA*-infected mice for 20 min, and only fully engorged mosquitoes were maintained for 10 d. To assess the number of oocysts that developed in the midgut, the mosquitoes were dissected at 10 d post-infection. The *TEP1*R1/~~S1m~~* (i.e. **R1* allele), **S2/~~S1m~~ (*S2)* and **S1/~~S1m~~ (*S1)*

alleles carried the lowest, intermediate and highest number of parasites respectively (**Fig. 3-2**A).

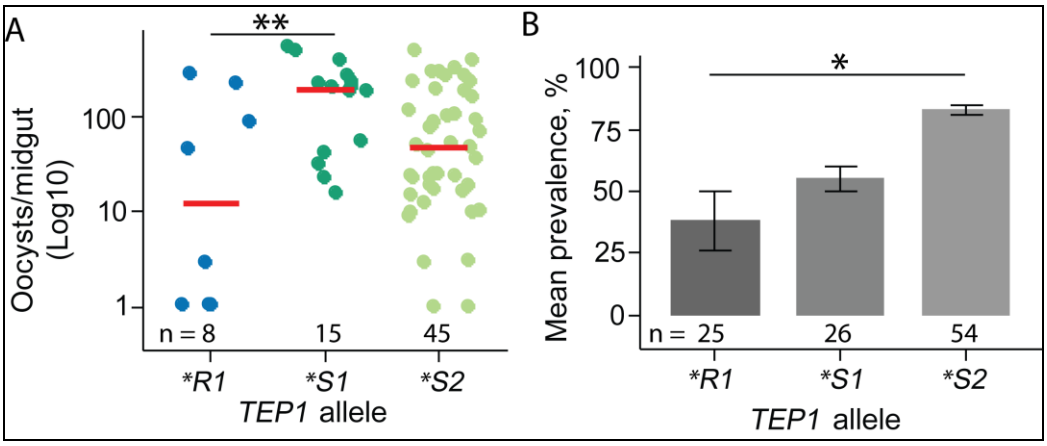


Fig. 3-2. Influence of *TEP1* alleles on *P. berghei* infection in *MH3T1* mosquitoes.
(A) *P.b GFPcon ANKA* strain infection loads in the *MH3T1* progeny in F1 generation. Each dot represents the number of oocysts per midgut. The red-colored line represents the median of the oocysts. Statistical significance between the groups was calculated using non-parametric Kruskal-Wallis with pairwise Wilcox test and Bonferroni correction. Results from two independent experiments were pooled. Statistically significant differences are indicated by $**p<0.01$.
(B) Corresponding percentage prevalence per each genotype group in the same F1 generation in Fig. 3-2A. Statistical significance between the groups was calculated using ANOVA test on normalized data, and Tukey's HSD test. Error bars show the mean ± SEM from the two independent experiments. Statistically significant differences are indicated by $*p<0.05$.

Interestingly, **S1* mosquitoes had higher oocyst loads than **S2* mosquitoes (**Fig. 3-2**A). In contrast, at the level of infection prevalence, more *S2 mosquitoes were infected than **S1* mosquitoes (**Fig. 3-2**B). However, in both cases the differences between **S1* and **S2* alleles were not statistically significant. The order of increasing parasite load; *TEP1*R1<*S2<*S1* was evident (**Fig. 3-2**A). I observed an increasing prevalence of infection (that is the proportion of infected mosquitoes relative to the total number of mosquitoes) in the following order *TEP1*R1<*S1<*S2* (**Fig. 3-2**B). The results obtained was similar to the published data on reciprocal allele-specific *RNAi* that linked *TEP1 *R1/R1* genotype to lower *P. berghei* development than *TEP1*S2* (5).

3.4.4 *TEP1*R1/R1* mosquitoes are more resistant to *Plasmodium* infections

In a control experiment, *H3T1* mosquito line was infected with *P. berghei* strain as previously described (5) (**Fig. 3-3**).

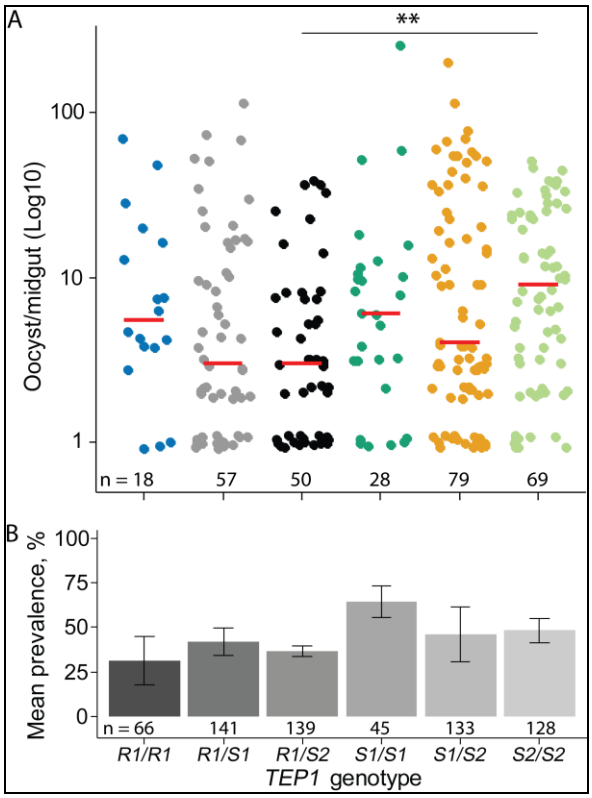


Fig. 3-3. Phenotype differences in *H3T1* mosquitoes upon *P. berghei* infection.
(A) *P. berghei* oocyst loads per midgut. Each dot represents the number of oocysts per midgut. The red lines represent the median of the oocysts. Statistical significant differences are indicated by $^{**}p<0.01$. Statistical significance between the groups was calculated using non-parametric Kruskal-Wallis with pairwise Wilcox test and Bonferroni correction. The numbers below the data points represent the sample size of the infected mosquitoes per genotype.
(B) *P. berghei* infection prevalence. The horizontal axes show *TEP1* genotype. The numbers below each bar represent the sample sizes per genotype. Results from 4 independent experiments were combined. Error bars show the mean ± SEM. Statistical significance between the groups was calculated using ANOVA test on normalized data, and Tukey's HSD test.

A statistically significant difference in the parasite oocyst loads was observed between **R1/S2* and **S2/S2* ($p<0.01$) (**Fig. 3-3**A). The *TEP1*R1/R1* genotype had the lowest prevalence of infection (27%), followed by **R1/S2* (46%), **R1/S1* (40%), **S2/S2* (54%), **S1/S2* (59%) and **S1/S1* (62%) in order of increasing susceptibility.

Similar infection experiments with human malaria parasite, *P. falciparum (NF54)* strain were performed on the same *H3T1* mosquito lines (**Fig. 3-4**).

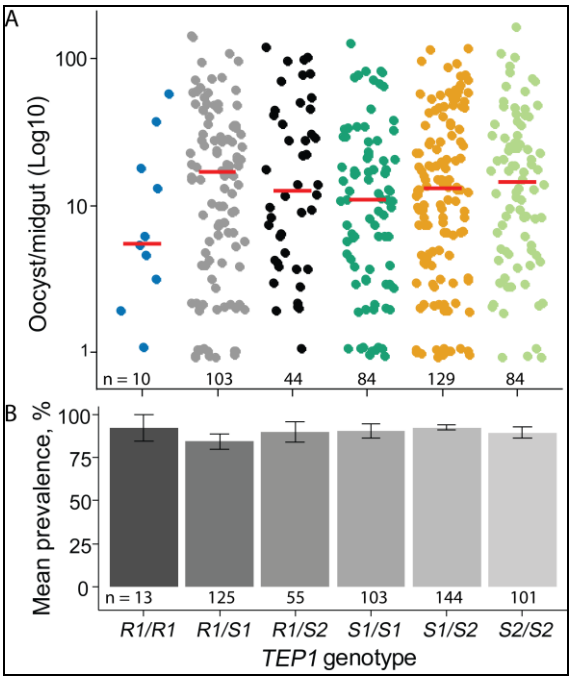


Fig. 3-4. Phenotype differences in *H3T1* mosquitoes upon *P. falciparum* infection.
(A) *P. falciparum* oocyst loads per midgut. Each dot represents the number of oocysts per midgut. The red bars represent medians of oocysts per midgut. Statistical significance between the groups was calculated from non-parametric Kruskal-Wallis with pairwise Wilcox test and Bonferroni correction. Results of 5 independent experiments were pooled. The numbers below the data points represent the sample size of the infected mosquitoes per genotype.
(B) *P. falciparum* prevalence of infection. Results of 5 independent experiments were pooled. The numbers below each bar represent the total sample sizes for each genotype. Error bars show the mean ± SEM. Statistical significance between the groups was calculated using ANOVA test on normalized data, and Tukey's HSD test.

A marked decrease in oocyst loads in the **R1/R1* genotype was not statistically significant as compared with the parasite loads of the *TEP1* genotype groups, probably because of the small sample size (**Fig. 3-4**A). All other *TEP1* genotypes were equally infected. In addition, the prevalence of the infection was considerably high (>80%) for all *TEP1* genotypes (**Fig. 3-4**B) and significantly higher than in infections with *P. berghei*. As these results show no statistically significant differences in the NF54 *P. falciparum* oocyst loads or prevalence of infection between the *TEP1* genotypes, it suggests that the NF54 *P. falciparum* isolate was largely resistant to TEP1-mediated killing in *H3T1* mosquitoes. To assess the same phenotypes in *P. falciparum* isolates that have been associated with a high degree of susceptibility to TEP1-mediated immune responses, I established cultures of TEP1-susceptible *P. falciparum* isolates.

3.4.5 Establishment of TEP1-sensitive *P. falciparum* cultures was unsuccessful

In order to evaluate the role of *TEP1* polymorphism in infection with different *P. falciparum* parasites, efforts were made to establish cultures of parasites known to be TEP1-sensitive; *HB3* (165), *7G8* (158, 167), *NF54 GFP-Luc* (166) carrying a disrupted *Pfs47* locus and *NF165* (104). To test the infectivity of the gametocyte cultures on the 3- to 5-d-old mosquitoes, mosquito strains of *H3T1, Mut6* (138), *TEP1 S1/S1* subset of Ngousso (168) and *7b* (78) were bred and infected separately with the parasites. The *7b* lines were used as positive *TEP1*-deficient mosquitoes. *P. falciparum NF54* strain was used as a positive control for infections.

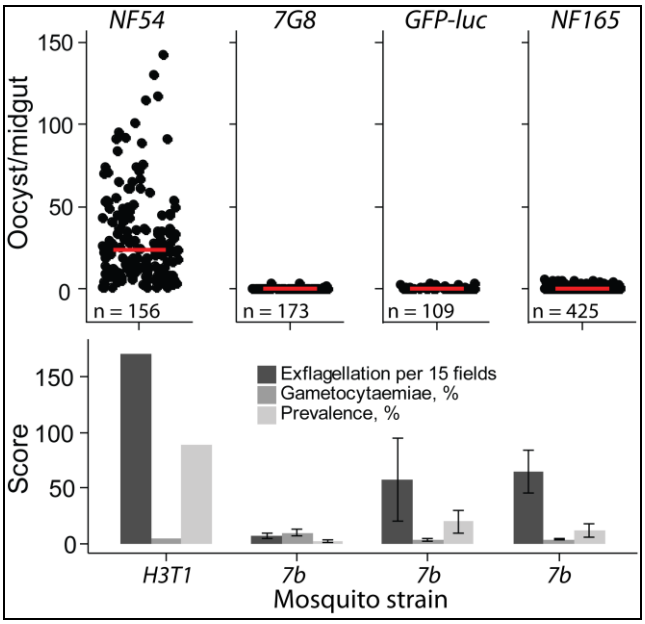


Fig. 3-5. Assessment of infectivity of TEP1-sensitive *P. falciparum* isolates.
Top panel shows *P. falciparum* oocyst loads per parasite strain to check whether they can infect mosquitoes. The red bars represent medians of oocysts per midgut. The numbers below the data points represent the sample size of the total mosquitoes that were dissected. Except for the *NF54,* multiple independent experiments were pooled. Bottom panel displays corresponding summary comparison of gametocytaemiae, exflagellation and prevalence in infections with *NF54*, *7G8*, *NF54HT-GFP-luc* and *NF165 P. falciparum* isolates in mosquitoes. In general, the *7G8*, *NF54HT-GFP-luc* and *NF165* parasites were not infectious to the *TEP1*-deficient mosquitoes; therefore, they were not be used for infection experiments with the wild-type *H3T1* mosquitoes.

.

To gauge the general performance of the cultures, profiles of three parameters; gametocytaemia, exflagellation/15 fields and infection prevalence were assessed in the infected mosquitoes. Establishment of the cultures of TEP1-sensitive parasites posed a number of challenges. First, repeated attempts to establish *HB3* gametocyte cultures failed to produce infectious gametocytes (data not shown). Second, the *7G8, NF54 GFP-Luc* and *NF165* parasites produced gametocytes, which did not generally infect the mosquitoes (**Fig. 3-5**). Although the *7G8* cultures showed high gametocytemia, repeated infections of the mosquito strains showed poor exflagellation and infection prevalence levels (**Fig. 3-5**). The *7G8* cultures were therefore discarded. Both *NF54 GFP-Luc* and *NF165* cultures showed good exflagellation levels but failed to infect mosquitoes. The *NF165* asexual cultures were highly susceptible to stress, such as slight changes in culture conditions. In addition, despite good gametocytaemia and high exflagellation rates, the *NF165* parasites were generally not infectious to mosquitoes. Therefore, the *NF165* cultures were also discarded. The difficulties in culturing these TEP1-susceptible parasite strains than the TEP1-resistant NF54, prevented us from evaluating the impact of *TEP1* alleles on *P. falciparum* development.

3.5 Discussion

Our understanding of functional influence of *TEP1* polymorphism on phenotypic traits during malaria infections provides important progress towards development control strategies of malaria transmission (*4*, *5*, *35*, *77*, *78*). These studies are facilitated by breeding and maintenance of the *A. gambiae* mosquito colonies, of which *TEP1* susceptible mosquitoes are easier to breed while refractory ones may be challenging (*35*). In this thesis, difficulties with the establishment of *A. gambiae H3T1* strain with equal frequencies of *TEP1*R1* or **S1* or **S2* alleles suggested high lethality rates of **R1/R1* genotypes. This study did not determine the optimal conditions and factors that benefit the survival of **R1/R1* homozygotes. Interestingly, similar difficulties were encountered with rearing mosquitoes with *TEP1*R2* allele, as the introgression of **R2* allele into *H3T1* strain was unsuccessful. Selection strategies were effective in increasing the frequencies of **R1* alleles but did not significantly improve the proportion of **R1* homozygotes. Nevertheless, modest numbers of **R1/R1* mosquitoes offered an opportunity to perform *P. berghei* and *P. falciparum* infections in *H3T1* and *MH3T1* lines.

Blandin *et al.* (5) compared resistance of *TEP1*R1/R1, *R2/R2* and **S2/S2* mosquitoes to *P. berghei* parasites, but did not examine the phenotype of **S1/S1* mosquitoes. Our results of *Plasmodium* infections of *MH3T1* indicated lower infection rates of the *TEP1*R1* hemizygotes as compared to **S2* and **S1*, where **S1* mosquitoes had higher oocyst numbers than **S2*. In contrast, at the level of infection prevalence, more **S2* mosquitoes were infected than **S1*. But since I could perform only two independent experiments, differences between **S1* and **S2* were not statistically significant. Therefore, I concluded that *TEP1*S* alleles are equally susceptible to *P. berghei* infections. These results are in line with the previous reports that showed that *TEP1*R1* is more resistant to *P. berghei* infections than the *TEP1*S2* (5).

Low numbers of **R1/R1* mosquitoes reduced the significance of our results with *P. berghei* infections, though I observed a trend towards higher refractoriness of **R1/R1* mosquitoes, whereas **S1/S1* mosquitoes showed higher susceptibility as compared to **S2/S2* homozygotes and other heterozygotes with intermediate phenotypes. Interestingly, *P. berghei* infections yielded lower prevalence than infections with *P. falciparum.* These observations are consistent with previous report that the *P. falciparum* NF54 used in this study is resistant to TEP1 (159). Recent reports suggested that some African parasites, such as *NF54* strain, have developed strategies to evade TEP1-mediated immune responses (158, 160). Unfortunately, I was unable to assess the impact of *TEP1* polymorphism on the development of TEP1-sensitive *P. falciparum* parasites due to difficulties with the maintenance of *in vitro* cultures.

3.6 Conclusion

My results support the previous reports demonstrating TEP1-mediated mosquito resistance to rodent malaria parasites and extend these observations to **S1* allele. Unexpectedly, the challenges associated with the establishment of mosquito line with equal representation of *TEP1* alleles suggest higher lethality rates of **R1* homozygote mosquitoes in the conditions of insectary. However, the observation that **R1* homozygotes are only found exclusively in *A. coluzzii* species in West Africa region (Chapter 2) suggests that the maintenance of this allele is driven by unknown environmental and *A. coluzzii*-specific conditions unique to specific locations in West Africa.

Chapter 4

General Discussion

General Discussion

4.1 Summary

The aim of this thesis was to: 1) elucidate how *TEP1* locus contributes to genetic structure of local *A. gambiae s.l.* species across four sub-Saharan Africa countries (Chapter 2); and 2) assess the impact of *TEP1* alleles and genotypes on *P. berghei* and *P. falciparum* development (Chapter 3). The main conclusions of my study are that: i) the *TED* region is sufficient for *TEP1* genotyping, including clear identification of *TEP1*S1* and **S2* alleles; ii) the high throughput PCR-RFLP genotyping approach offers a robust and affordable strategy for both small and large scale *TEP1* genotyping; iii) a novel *TEP1*R3* allele was identified in *A. merus* population in coastal Kenya; iv) ecotype partitioning of local malaria vectors across Africa was described using *TEP1* genotyping; v) a 'conditional lethality' of *TEP1*R1* allele in **R1* individuals, occurs under the standard laboratory breeding conditions suggesting specific natural factors that promote fixation of the allele in *A. coluzzii* in West Africa; and vi) *TEP1*S1* and *TEP1*S2* alleles are equally susceptible to malaria parasite infections.

4.2 *TED* region identifies all the *TEP1* allele subclasses

The *TED* region offers a genetic marker for *TEP1* genotyping as it can be used to identify all the *TEP1* alleles with the ability to provide clear distinction between *TEP1*S1* and **S2* alleles, which were previously unresolved (*6*, *8*, *35*). The developed method is also affordable for *TEP1* genotyping in the field laboratories. By combining species-specific markers with *TEP1* genotyping, we identified partitioning of local malaria vectors into ecotypes across African environments. Since *TEP1* locus shows clear geographic and species-specific patterns, it can be potentially used to monitor the on-going speciation events caused by climate change, dispersal and colonization of new ecological niches, local adaptation, and gene flow between vector species.

4.3 Natural selection drives biogeographic genetic diversity at *TEP1* locus

Natural selection drives the global distribution of species and of *TEP1* alleles across Africa (*6*, *8*, *35*). Local (habitats) selection forces which form the hallmark of local adaptation in *A. gambiae* mosquitoes, allow the mosquitoes to accumulate local alleles that make them adapt to local pathogen and environments (*32*, *129*, *132*, *154*). This study has biogeographically categorized the *TEP1* alleles and genotypes into specialist and generalist groups (**Fig. 4-1**).

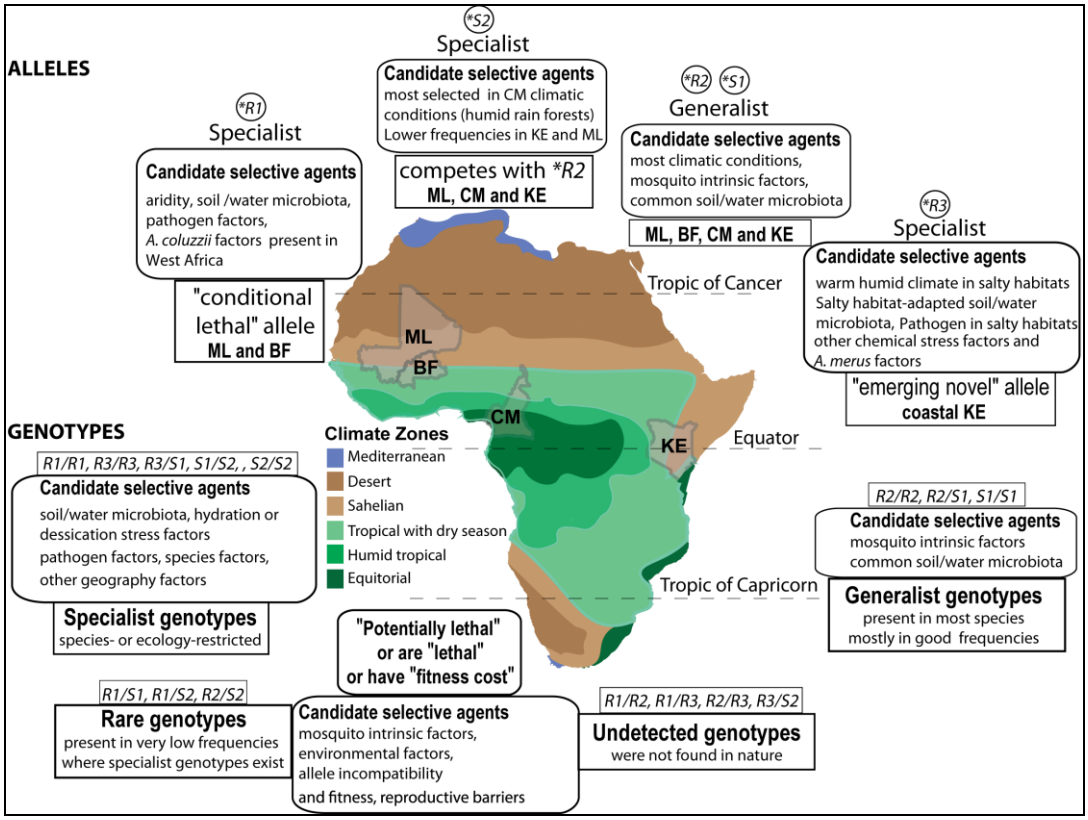


Fig. 4-1. Hypothesis underlying natural forces acting on *TEP1* locus.
The *TEP1* alleles and genotypes were categorized according to their geographical and species distribution; generalist and specialist alleles. This study hypothesizes that prime selective agents are specific to their ecological biotopes and species, and highlights the most likely agents for selections as a proxy for future field studies seeking to identify actual potent selective agents.

In Africa, *TEP1*R1/R1* genotypes inhabit sahelian ecosystem, arid zones and in Mali and Burkina Faso (8, 35). This thesis upholds evidence from these previous reports that *TEP1*R1* alleles are locally selected in West Africa. Unknown environmental and species-specific conditions prevailing only in West Africa drive enrichment of *TEP1*R1* allele (or *TEP1*R1/R1* genotype) in *A. coluzzii* at near-fixation (6, 8, 35). In addition, our study uncovered high lethality of homozygous **R1/R1* mosquitoes under laboratory conditions. Therefore, from the challenges of breeding the **R1/R1* mosquitoes in the laboratory (35) and in our study, I propose that the **R1* allele is a conditional lethal. In homozygote individuals, a conditional lethal allele may be maintained by some special unknown circumstances without which they selected against or are rarely selected for and/or is permitted more at heterozygous (**R1/S1* and **R1/S2*) genotypes.

Our study has extended this observation on local selection to *TEP1*R3* alleles among *A. merus* population in coastal Kenya, and **S2* alleles mainly in Cameroon by *A. coluzzii* and *A. gambiae s.s.*. As the *TEP1*R3* allele portrays unique polymorphism, it

suggests that these modifications may be particularly important for certain fitness of these species in harsh salty-water environment. Furthermore, it may reflect the impact of evolution and natural selection on *TEP1* locus towards local adaptation. Further, *TEP1*S1* and **R2* alleles, and genotypes they form, have a generalist nature, where they are selected almost in all sites and species. Of note, the **R2/S1* genotype (with the exception of *A. coluzzii* in Mali and Vale de Kou in Burkina Faso) is maintained in all the studied species across Africa suggesting that it competitively outperforms other heterozygotes, however, selection forces that benefit this genotype and **R2/*S1* allele compatibility remain unknown.

Importantly, *P. berghei* infection experiments showed that both **R2* and **S2* alleles render the mosquitoes susceptible to infections as compared to **R1* alleles (5). The cooler and humid climate prevailing in Cameroon apprears to promote selection for the specialist **S2* alleles. Here, our data have shown further that **S2* and **S1* mosquitoes are equally susceptible to both *P. berghei* and *P. falciparum* infections, suggesting that these alleles confer higher vector competence to *A. gambiae* in local populations, hence have a direct consequence of malaria transmission.

4.4 Open questions and future directions

Our study opens new research avenues (**Fig. 4-1**) that should address the following questions;

First, what determines restricted selection patterns of *TEP1*R1* and **S2* mosquitoes across African biotopes? What drives the 'conditional lethality' of the *TEP1*R1* allele?

Second, the structured genotype diversity revealed in this study, in particular the enrichment of *A. coluzzii* **R1/R1* genotypes in Mali and Burkina Faso, **R2/S1* across Africa, **S1/S2* in Cameroon, and **R3/R3* and **R3/S1* coastal Kenya needs further investigation to establish their biological regimes and functional constraints underlying this polymorphism.

Third, could there be natural beneficial roles or selective advantages of **R2/S1,* **R3/S1* and **S1/S2* heterozygotes over homozygote counterparts resulting in traits that influence malaria transmission?

Fourth, future studies should examine the role of the **R3* allele in *A. merus* mosquito resistance to *Plasmodium*, or identify its role in other biological processes or whether it has a risen as a result of local adaptation to salty water biotopes in comparison to *A. melas* counterpart in West Africa. Is the **R3* allele undergoing

fixation in *A. merus* larval populations? Genotyping of *TEP1* in a cross sectional larval collections spanning dry and rainy seasons may be a good starting point. Towards this end, a short preliminary fieldwork was conducted in the same sampling sites in Malindi, coastal Kenya, where the **R3* allele was identified. Larval populations from two discrete larval habitats have been collected (data not shown) and *TEP1* analyses are underway. These data are expected to inform the design and the direction of a bigger study.

Fifth, sequence similarity in SNPs and amino acid substitutions between **R3* allele and **R2/*S* alleles suggest novelty in derivation of the **R3* allele, and point towards inter-allele genetic exchange resulting in extensive shared polymorphism. Whether there are high mutational constrains (salinity and water chemistry or nature of predators, pathogens or vegetation) in the saline breeding environments, which may have triggered the emergence of this novel allele and its maintenance in the population, remains to be addressed (**Fig. 4-1**). In particular, the specific **R3* nucleotide and amino acid modifications discussed here, may potentially confer some fitness advantage to *A. merus* mosquito populating in these saline biotopes that could impact significantly on malaria epidemiology. To understand molecular events underlying this fitness, characterization of genetic recombination events or mutation pressures acting on the entire genome especially on the chromosome 3 in *A. merus* populations may provide some light. Interrogation of the published data on the sequences, the genome assembly and transcriptomes of *A. merus,* one of the 16 recently published genomes of *A. gambiae*, may offer an insightful basis (24).

And sixth, the challenge is to identify natural selection forces that cause the geographic- and species variation in *TEP1* polymorphism in the African mosquito populations.

4.5 Conclusion

In the *A. gambiae s.l.* populations, *TEP1* polymorphism has been maintained throughout Africa (6, 8, 35). This thesis has shown that polymorphism (*TEP1*R2* and **S1)* is substantially shared among species, while (*TEP1*R1, *R3* and **S2*) are locally selected by specific *Anopheles* species. Particularly, the *TEP1*R2* and **S1* are the most shared alleles across Africa. As they were found in all the four countries, it suggests that the species in the *A. gambiae s.l.* complex may have had these alleles before they expanded to colonize new ecological niches and speciate to sibling species.

The generalist **S1* allele suggests compatibility with all the other alleles as exemplified by presence of **R1/S1, *R2/S1, *R3/S1* and **S1/S2* heterozygotes, a phenomenon underpinning unknown functional fitness advantage driving this compatibility. This geographic variation in allele and genotype selections may be, in part, due to differences in the choice and stability of breeding habitats (*118*) or prevailing ecological climates of which local factors and geographical isolation significantly drive speciation, diversification and selection of fitter genotypes (*98*, *118*).

Collectively, the findings from this study suggest a trade-off between intrinsic (vectorial) forces and extrinsic (ecological) factors that drive the vector adaptation to local ecological ecotypes through strict maintenance of the *TEP1* polymorphism and microevolution (i.e. evolution in within the population leading to change in allele and genotype) at the habitat level. Additionally, it suggests local selection for specialist alleles is advanced by specialist species while global dispersal of generalist alleles is driven by generalist species. These data suggest that *TEP1* locus experiences distinct local selective pressures in mosquito natural populations, in agreement with the proposed pleiotropic functions of *TEP1* (*80*) i.e. influence of *TEP1* gene on more than one phenotypic traits such as in reproduction (*80*) and immune responses (*4*, *5*, *18*, *71*). Further, this study proposes a 'conditional lethality' of *TEP1*R1* allele in the laboratory conditions, which is promoted by unknown conditions; and demonstrates that *TEP1*S1* and *TEP1*S2* alleles are equally susceptible to malaria parasite infections.

The development of novel vector control measures requires in-depth documentation and prediction of demographic vis-à-vis ecological events underpinning local adaptation, speciation, extinction, colonization of new niches in malaria vectors, like those driven by the climate change and human activities (*48*, *108-111*). Using *TEP1* as a marker, revealed the partitioning of *A. gambiae s.l.* species into generalist and specialist genotypes. Therefore, it offers clear geographic and species-specific patterns, and can be potentially used to monitor the on-going speciation events dispersal and colonization of new ecological niches, local adaptation, and gene flow between vector species. Importantly, this study has contributed molecular tools for high throughput *TEP1* genotyping that can be used to complement species identification methods in surveying and monitoring the population dynamics of local malaria vectors over time and space. It has provided phylogeny and sequence information needed to improve the understanding on *TEP1* diversity. Incorporating these informative ecological genetic markers in

malaria control programs and in other vector-borne diseases should help in forecasting future demographic population dynamics that come with severe epidemiological consequences.

Appendices

Appendix 1. Materials, Equipment and Software used in this study

Materials, Equipment and Software		
A	**Biological Materials and where the Resources were obtained from.**	
No.:	**Material**	**Source**
1	*A coluzzii* (Ngousso colony) mosquito strains	Institut de Biologie Moléculaire et Cellulaire (IMBC), Strasbourg, France (*168*).
2	*A. coluzzii* L3-5 refractory mosquito trains	Dr. Stéphanie Blandin, IMBC, Strasbourg, France (*167*, *169*).
3	*G3* mosquito stains	Dr. Flaminia Catteruccia. Department of Immunology and Infectious Diseases, Harvard T.H. Chan School of Public Health, Boston, Massachusetts 02115, USA. 2 Center for Communicable Disease Dynamics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts 02115, USA.
4	4Arr mosquito strain	MR4 (*5*).
5	*7b TEP1* knockdown mosquito strain	IMBC, Strasbourg, France (*78*).
6	*Mut6* - *TEP1* mutant mosquito strain	Dr. Erick Marois, IMBC, Strasbourg, France (*138*).
7	*P. berghei* ANKA	(*163*).
8	*P. falciparum* NF54	Prof. Dr. Robert W. Sauerwein at Department of Medical Microbiology, Radboud University Medical Centre, Nijmegen, The Netherlands (*104*, *164*).
9	*P. falciparum* NF54HT-GFP-luc malaria parasites	Prof. Dr. Robert W. Sauerwein *(166)*
10	*P. falciparum* HBC malaria parasites	MR4 *(165)*
11	*P. falciparum* NF165 malaria parasites	Prof. Dr. Robert W. Sauerwein (*104*)
12	*P. falciparum* 7G8 malaria parasites	MR4 (*158*, *167*)
13	DH5α and 10βeta chemically competent *E. coli* cells	New England Biolabs, UK

B	**Consumable Materials**	
No.:	**Material**	**Supplier (Manufacturer), Country**
14	1 kb DNA Ladder	Thermo Fisher Scientific, Germany
15	1 kb Plus DNA Ladder	Thermo Fisher Scientific, Germany
16	100 bp Plus DNA Ladder	Thermo Fisher Scientific, Germany
17	2-Propanol	Carl Roth, Germany
18	Agarose	Invitrogen, Germany

Materials, Equipment and Software		
No.:	**Material**	**Supplier (Manufacturer), Country**
19	Ammonium Acetate	Ambion, Germany
20	Ampicillin	Sigma-Aldrich, Germany
21	Ampicillin (stock solution 1000×)	100 mg/ml in ddH2O, sterile filtered
22	Bacto-Agar	Invitrogen, Germany
23	Calcium Chloride	Merck, Germany
24	Coffee-Cup Exclusiv 120 ml and 300mL	IGEFA, Germany
25	Coverslips	Carl Roth, Germany
26	CryoTubes, 1.8 ml (cryopreservation of bacteria)	Greiner Bio-One, Germany
27	Dimethyl Sulphoxide (DMSO)	Sigma-Aldrich, Germany
28	dNTP-mix	Thermo Fisher Scientific, Germany
29	Electrophoresis buffer TAE 50 ×	Thermo Fisher Scientific, Germany
30	Electrophoresis buffer TBE 50 ×	Thermo Fisher Scientific, Germany
31	Eppendorf tubes, 1.5 ml and 2 ml	Eppendorf, Germany
32	Ethanol, molecular grade	Carl Roth, Germany
33	Ethidium Bromide	Sigma-Aldrich, Germany
34	Ethylenediamine tetraacetic acid (EDTA)	Carl Roth, Germany
35	Falcon tubes, 15 ml and 50 ml	Sarstedt, Germany
36	Gel Loading Dye, Blue (6 ×)	Thermo Fisher Scientific, Germany
37	Gentamycin	Sigma-Aldrich, Germany
38	Giemsa concentrate (10 ×)	VWR, Germany
39	Giemsa staining buffer (1 ×)	VWR, Germany
40	Giemsa staining solution (1 ×)	Giemsa staining buffer, Giemsa concentrate
41	Glass beads	Sigma-Aldrich, Germany
42	Glycerol	Carl Roth, Germany
43	Golden Fish	Tetra, Germany
44	Human A+ serum (pool from > 20 donors)	HAEMA, Blood Bank, Germany
45	Human red blood cells (O positive, pool from > 8 donors)	HAEMA, Blood Bank, Germany
46	Hypoxantine liquid	Neustadt, Germany
47	Inoculation loops, plastic, disposable	VWR International, Germany
48	Kanamycin	Sigma-Aldrich, Germany
49	Ketamine	Arzneimittelvertrieb, Germany
50	LB-Agar	Invitrogen, Germany
51	Luria Broth (LB) base	Invitrogen, Germany
52	Magnesium Chloride	Merck, Germany
53	Magnesium Sulphate	Merck, Germany
54	MicroAmp Fast Optical 96-Well Reaction Plates	Applied Biosystems, Germany
55	MicroAmp Optical Adhesive Film	Applied Biosystems, Germany
56	Microscope slides	Menzel, Germany

Materials, Equipment and Software		
No.:	**Material**	**Supplier (Manufacturer), Country**
57	*P. falciparum* complete medium (asexual cultures) (RPMI 1641, 10% Human A+ serum, 2% hypoxantine liquid, 20 µg/ml gentamycin, sterile filtered (0.22 µm) and stored at 4 °C for max. 3 weeks, or -20 °C for long term storage)	Gibco Invitrogen, Germany
58	*P. falciparum* complete medium (gametocyte cultures) (RPMI 1641, 10% Human A+ serum, 2% hypoxantine liquid, sterile filtered (0.22 µm) and stored at 4 °C for max. 2 weeks, or -20 °C for long term storage)	Gibco Invitrogen, Germany
59	Parafilm	Bemis, USA
60	Pasteur pipettes	Carl Roth, Germany
61	PBS, sterile solution	Gibco Invitrogen, Germany
62	PCR tubes	Greiner Bio-One, Germany
63	Petri dishes	Greiner, Germany
64	pGEM-T Easy	Promega, USA
65	Phosphate-buffered saline (PBS), tablets	Gibco Invitrogen, Germany
66	Phusion High-Fidelity DNA Polymerase	New England Biolabs, Germany
67	Pipette tips, 10-1000 µl	Sarstedt, Germany
68	Proteinase K	Invitrogen, Germany
69	QIAprep Spin Miniprep Kit	Qiagen, Germany
70	QIAquick PCR Purification Kit	Qiagen, Germany
71	Restriction enzyme, FastDigest kit (*Bam*HI, *Hin*dIII, *Bse*NI and *Nco*I)	Thermo Fisher Scientific, Germany And Fermentas, USA
72	Restriction enzymes kit (*Bam*HI, *Hin*dIII, *Bse*NI and *Nco*I)	New England, UK
73	RPMI 1641, with L-glutamine and 25mM HEPES	Gibco Invitrogen, Germany
74	Salt, table salt	LIDL, Germany
75	Sea salt	Alnatura, Germany
76	Super Optimal Culture (SOC) medium	Thermo Scientific, Germany
77	Sodium Acetate	Carl Roth, Germany
78	Sodium Chloride	Carl Roth, Germany
79	Sodium dodecyl sulphate (SDS)	Sigma-Aldrich, Germany
80	Sugar, table sugar	LIDL, Germany
81	Go *Taq* DNA polymerase kit	Promega, USA
82	Tris-Base	Carl Roth, Germany
83	Triton X-100	Carl Roth, Germany
84	Tween-20	Sigma-Aldrich, Germany

Materials, Equipment and Software		
No.:	**Material**	**Supplier (Manufacturer), Country**
85	Xylazine	Bayer Vital, Germany

C	**Equipment**	
No.:	**Material**	**Supplier (Manufacturer), place**
86	Aspirator, mechanical	Clarke, USA
87	Bunsenburner LABOGAZ 470	Carl Roth, Germany
88	Centrifuge 5804, for bacterial cultures	Eppendorf, Germany
89	Centrifuge, 96-well plates, benchtop	Thermo Fisher Scientific, Germany
90	Desktop office computer	DELL, USA
91	Fragment Analyser	Advance Analytical, USA
92	Freezer -20 °C	Liebherr, Germany
93	Freezer -80 °C	Liebherr, Germany
94	Fridge 4 °C	Siemens, Germany
95	Gel casting system	VWR International, Germany
96	Gel documentation system Gel Doc 2000	BioRad, Germany
97	Glass Midi-feeder 1.5 ml (for mosquito blood feeding)	Coelen Glas, Germany
98	Glassware	Schott, Germany
99	Ice machine AF200	Scotsman, Germany
100	Incubator and shaker Innova 40 (for liquid bacterial cultures)	Eppendorf, Germany
101	Incubator freezer BK 160 (for mosquitoes)	Thermo Fisher Scientific, Germany
102	Incubator Heraeus B6 (for plated bacterial cultures)	Thermo Scientific, Germany
103	Ipad	Apple, USA
104	Laminar flow, Herasfe KS12	Thermo Scientific, Germany
105	MacBook Air laptop	Apple, USA
106	Microscope Leica DM2500	Leica, Germany
107	Microscope Zeiss Axio Observer.Z1	Zeiss, Germany
108	Microscope Zeiss Stemi 2000-C Stereo	Zeiss, Germany
109	Mosquito cages	Bugdorm, USA
110	Nanodrop 2000c	Thermo Scientific, Germany
111	PCR Thermocycler MJ Mini	Bio-Rad, Germany
112	Pipette (2.5 µl , 20 µl , 200 µl, and 1000 µl)	Eppendorf, Germany
113	Pipetting aid Pipetus	Hirschmann Laborgeräte, Germany
114	Portable Timer	Carl Roth, Germany
115	Sterile hood HERAsafe KS	Thermo Scientific, Germany
116	Thermo-mixer Mixmate	Eppendorf, Germany
117	TissueLyser LT	Qiagen, Germany
118	Vortex	Vortex Genie, USA/ Scientific Industries, USA

Materials, Equipment and Software		
No.:	**Material**	**Supplier (Manufacturer), place**
119	Water bath, GFL 1002	GFL, Germany

D	**Online Software and Software and database**		
No.:	**Online server or software**	**Purpose**	**Link or Reference**
120	NCBI BLAST	BLAST, Sequence retrival and alignment	NCBI, USA
121	Bioinformatics.org	Primer design , Sequence *in silico* manipulations	www.bioinformatics.org/sms2
122	Justbio	Primer design , Sequence *in silico* manipulations	www.justbio.com
123	Datamonkey server	exploring sequence alignments for evidence of selection forces acting the genes	http://www.datamonkey.org/
124	Kalign	Full-length multiple sequence alignments	http://www.ebi.ac.uk/Tools/msa/
125	Boxshade	Full-length multiple sequence alignments	http://www.ch.embnet.org/software/BOX_form.html
126	SNAP	Calculate codon-based cumulative synonymous and non-synonymous substitutions	references (*151*, *152*)
127	MEGA6	Phylogenetic sequence analyses	Reference (*148*)
128	PAML	Phylogenetic sequence analyses and exploring sequence alignments for evidence of selection forces acting the genes	Reference (*170*, *171*)
129	BEAST	Phylogenetic sequence analyses	Reference (*150*)
130	DnaSP version 5.0	Sliding window analyses and neutrality tests; Tajima's D, Fu and Li's D and Fu and Li's F statistics	Reference (*142*, *143*)
131	ProSize software version 2.0	Analyze PCR and RFLP fragments	Advance Analytical Technical, USA
132	TCS1.21	Estimating gene genealogies	Reference (*144*)
133	Pymol	Molecular visualization and manipulation of TEP1 crystal structure	www. pymol.org
134	PhyML 3.0	Phylogenetic sequence analyses	Reference (*149*)
135	NCBI Primer-Blast	Primer design	NCBI, USA
136	NCBI PubMed	Search engine for scientific publications	NCBI, USA

Materials, Equipment and Software			
137	VectorBase	Bioinformatics resource for invertebrate vectors of human pathogens	NIAID, USA
138	Adobe Photoshop CS5	Image editing	Adobe Systems Inc., USA
No.:	**Software/online server**	**Purpose**	**Link or Reference**
139	Adobe Illustrator CS5	Figure preparation	Adobe Systems Inc., USA
140	MS Office	Preparation of Word /PowerPoint slides	Microsoft, USA
141	EndNote	Bibliography Reference Manager	ENDNOTE, Netherlands
142	SeqMan Pro	Sanger sequence assembly	DNASTAR, USA
143	Bioedit	Editing, aligning sequences	T. A. Hall *(139)*
144	qGIS	GIS, preparation of maps	http://www.qgis.org/en/site/
145	FACS	Checking parasitaemia in mice	LSR Fortessa, USA
146	R (version 3.3.2)	Data analysis, statistics, and visualization	R Core Team (2016) (*141*)
147	RStudio (version 0.99.893)	Data analysis, statistics, and visualization	RStudio Team (2015)
148	r genetic package (version 1.3.8.1)	Analyse population genetics	R Repository (CRAN)
149	dplyr R package (version 0.5.4)	Data manipulation in R	R Repository (CRAN)
150	plyr R package (version 1.8.4)	Splitting, applying and combining data in R	R Repository (CRAN)
151	ggplot2 R package (version 2.2.1)	Data visualizations in R	R Repository (CRAN)
152	cowplot R package (version 0.7.0)	Streamline plot theme and plot annotations in 'ggplot2'	R Repository (CRAN)
153	reshape2 R package (version 1.4.2)	Reshape data	R Repository (CRAN)
154	gridExtra R package (version 2.2.1)	Apply functions for grid graphics	R Repository (CRAN)

Appendix 2. Sample R scripts used to visualize the distribution of *TEP1* variation

Global and local distribution of *TEP1* genotypes and alleles in Africa

A Data description

```
#R-Script to analyse the global and local levels (i.e. per country and sampling sites resüpectively) TEP1 genotypes. I stratified these TEP1 genotypes and allelic frequencies according to species, countries and sampling sites in Mali, Burkina Faso, Cameroon and Kenya, and specified color-blind friendly colors to be used in the graphs.
cbbPalette <- c("#0072B2", "#999999", "#000000", "#56B4E9", "#D55E00","#9999CC", "#CC79A7", "#F0E442", "#009E73", "#E69F00", "#B2DF8A")
#Respectively these colors were assigned to the following TEP1 genotypes and alleles: #1. "R1/R1" and "R1", #2. "R1/S1", #3. "R1/S2", #4. "R2/R2" and "R2", #5. "R2/S1", #6. "R2/S2", #7. "R3/R3" and "R3", #8. "R3/S1", #9. "S1/S1" and "S1"#10. "S1/S2", #11. "S2/S2"and "S2".
#set default working directory (wdir) accordingly, where data files for analyses are stored and are fetched by R from. Put the data files in the wdir.
getwd ()#check the current wdir
rm(list = ls())#clear environment
#MALVECBLOK data from our local data base (db)
db <- read.csv ("mvb.1.csv",sep=",",dec=".")#load packages
#library(genetics)# my known issue with the genetic R package is that it confuses the colors pellete assigned to the genotypes as some of its functions overlap/mask with functions from other loaded packages above. So I load it when I really need it. For now, I deactivate it.
my_packages <- c("plyr", "ggplot2", " reshape2", " gridExtra", "cowplot")
libraries(my_packages)
#column names are "Individual.ID", "Country", "Sample.ID", "Site.Name", "SpeciesAbbr", "Species", "Genotype"
#Abbreviation of countries by inserting a new column "CountryAbbr"
db$CountryAbbr[as.factor(db$Country) == "Mali"] <- "ML"
db$CountryAbbr[as.factor(db$Country) == "Burkina Faso"] <- "BF"
db$CountryAbbr[as.factor(db$Country) == "Cameroon"] <- "CM"
db$CountryAbbr[as.factor(db$Country) == "Kenya"] <- "KE"
```

B Global *TEP1* genotype distribution in Africa

```
#Overview of TEP1 genotypes across Africa
db1 <-subset(db, !SpeciesAbbr == "MS")#Omit the hybrids “MS”
spp<-ggplot(db1, aes(x=factor(CountryAbbr, levels = c("ML", "BF", "CM", "KE")),fill=factor(Genotype)))+geom_bar(position="fill", width=0.62)+facet_wrap(~Species)
spp<- spp+ggtitle ("Overview of TEP1 genotypes per species") +
  xlab ("Country") + ylab("Genotype frequency") +
  theme(legend.title = element_text(colour="black", size=10, face="bold")) +
  theme(legend.text = element_text(colour="black", size=10, face="italic"))
spp<-spp+ theme(legend.position="right") +scale_fill_manual(values=c("#0072B2", "#999999", "#000000", "#56B4E9", "#D55E00","#9999CC", "#CC79A7", "#F0E442", "#009E73", "#E69F00", "#B2DF8A"),name="Genotype",labels=c("R1/R1", "R1/S1", "R1/S2", "R2/R2", "R2/S1", "R2/S2","R3/R3", "R3/S1", "S1/S1", "S1/S2","S2/S2"))
spp#Fig. 2-9 Overview of global TEP1 genotype distribution across Africa
```

Global and local distribution of *TEP1* genotypes and alleles in Africa

C Local *TEP1* genotype distribution

```
#Abbreviation of countries by inserting a new column "SiteAbbr2" to abbreviate the long sampling sites
db$SiteAbbr2[as.factor(db$Site.Name) == "Nankilabougou"] <- "NK (Nankilabougou)"
db$SiteAbbr2[as.factor(db$Site.Name) == "vk5"] <- "VK5 (Vale de Kou 5)"
db$SiteAbbr2[as.factor(db$Site.Name) == "vk7"] <- "VK7 (Vale de Kou 7)"
db$SiteAbbr2[as.factor(db$Site.Name) == "somousso"] <- "SM (Somousso)"
db$SiteAbbr2[as.factor(db$Site.Name) == "Nkolbisson"] <- "NS (Nkolbisson)"
db$SiteAbbr2[as.factor(db$Site.Name) == "Nkolkoumou"] <- "NM (Nkolkoumou)"
db$SiteAbbr2[as.factor(db$Site.Name) == "Nkolondom"] <- "ND (Nkolondom)"
db$SiteAbbr2[as.factor(db$Site.Name) == "Mfou"] <- "MF (Mfou)"
db$SiteAbbr2[as.factor(db$Site.Name) == "Mvan"] <- "MV (Mvan)"
db$SiteAbbr2[as.factor(db$Site.Name) == "Ahero"] <- "AH (Ahero)"
db$SiteAbbr2[as.factor(db$Site.Name) == "Kakamega"] <- "KK (Kakamega)"
db$SiteAbbr2[as.factor(db$Site.Name) == "Busia/Teso"] <- "BT (Busia Teso)"
db$SiteAbbr2[as.factor(db$Site.Name) == "KWALE"] <- "KW (Kwale)"
db$SiteAbbr2[as.factor(db$Site.Name) == "KILIFI"] <- "KL (Kilifi)"
db$SiteAbbr2[as.factor(db$Site.Name) == "MALINDI"] <- "MD (Malindi)"
#Abbreviation of species by inserting a new column "SpeciesAbbr2" and filling with respective
#species abbreviations
db$SpeciesAbbr2[as.factor(db$SpeciesAbbr) == "M"] <- "A. col"
db$SpeciesAbbr2[as.factor(db$Species) == "A. gambiae s.s."] <- "A. gam"
db$SpeciesAbbr2[as.factor(db$Species) == "MS"] <- "A.col/gam"
db$SpeciesAbbr2[as.factor(db$Species) == "A. arabiensis"] <- "A. ara"
db$SpeciesAbbr2[as.factor(db$Species) == "A. merus"] <- "A. mer"
# ML: One site Nankilabogou site (NK)
#In BF: Three ecological villages sites Vale de Kou (VK5 and VK7) and Somousso (SM)
#combined graph of ML and BF with respect to species: A. coluzzii and A. gambiae s.s. in their respective sampling sites. I organize by site
mlbf<-subset(db, CountryAbbr!= "CM")
mlbf<-subset(mlbf, CountryAbbr!= "KE")
#mlbf<-subset(mlbf, SpeciesAbbr!= "MS")
#re-order the factors
mlbf$SiteAbbr2_f = factor(mlbf$SiteAbbr2, levels=c('NK (Nankilabougou)','VK5 (Vale de Kou 5)', 'SM (Somousso)','VK7 (Vale de Kou 7)'))
#ML and BF#
MLBFsite<-ggplot(mlbf, aes(x=factor(SpeciesAbbr2),fill=factor(Genotype))) +
  geom_bar(position="fill", width =0.5)+ facet_wrap(~SiteAbbr2_f)
MLBFsite<- MLBFsite+scale_fill_manual(values=c("#0072B2", "#999999", "#000000",
    "#56B4E9", "#D55E00", "#009E73"),name="Genotype",labels=c("R1/R1", "R1/S1",
"R1/S2", "R2/R2", "R2/S1", "S1/S1"))
MLBFsite<-MLBFsite+ggtitle ("Sympatric and allopatric vector populations in ML and BF") +xlab ("Species") + ylab("Genotype frequency") +theme(legend.title =
element_text(colour="black", size=10, face="bold")) +theme(legend.text =
element_text(colour="black", size=10, face="italic"))
#MLBFsite<-MLBFsite+ theme(legend.position="right")
MLBFsite#Fig 2-10A
```

Global and local distribution of *TEP1* genotypes and alleles in Africa

```
cm$SiteAbbr2_f = factor(cm$SiteAbbr2, levels=c('MF (Mfou)','MV (Mvan)', 'NS (Nkolbisson)', 'ND (Nkolondom)','NM (Nkolkoumou)')) # order the factor (sites) in CM
CMsite<-ggplot(cm, aes(x=factor(SpeciesAbbr2),fill=factor(Genotype)))
+geom_bar(position="fill", width=0.62)+ facet_wrap(~SiteAbbr2_f)
CMsite<-CMsite+ggtitle ("Different vector populations in CM select for similar genotypes") + xlab ("Species") + ylab("Genotype frequency") +
  theme(legend.title = element_text(colour="black", size=10, face="bold")) +
  theme(legend.text = element_text(colour="black", size=10, face="italic"))
CMsite<-CMsite+ theme(legend.position="right")
CMsite<- CMsite+scale_fill_manual(values=c("#56B4E9", "#D55E00","#9999CC", "#009E73", "#E69F00","#B2DF8A"),name="Genotype",labels=c("R2/R2", "R2/S1", "R2/S2", "S1/S1","S1/S2","S2/S2"))
CMsite#Fig 2-10B
ke<-subset(db, CountryAbbr=="KE")#KE
KEsite<-ggplot(ke, aes(x=factor(SpeciesAbbr2),fill=factor(Genotype)))
+geom_bar(position="fill", width=0.62)+ facet_wrap(~SiteAbbr2)
KEsite<-KEsite+scale_fill_manual(values=c( "#56B4E9", "#D55E00","#9999CC", "#CC79A7", "#F0E442", "#009E73", "#B2DF8A"),name="Genotype",labels=c("R2/R2", "R2/S1", "R2/S2","R3/R3", "R3/S1", "S1/S1","S2/S2"))
KEsite<-KEsite+ggtitle ("Different vector populations in KE select for similar genotypes") + xlab ("Species") + ylab("Genotype frequency") +
  theme(legend.title = element_text(colour="black", size=10, face="bold")) +
  theme(legend.text = element_text(colour="black", size=10, face="italic"))
KEsite<-KEsite+ theme(legend.position="right")
KEsite#Fig. 2-10C
```

D Local *TEP1* allele frequency distribution

```
adb <- read.csv ("mvb.1.csv",sep=",",dec=".")
#merging sites into sampling regions in BF (VK5 and VK7, and SM), CM (All sites together as CM), and KE (into western Kenya (WK) and coastal Kenya (CK)) into the column "MergedAbbr". Mali-Nankilabougou remains to be NK
adb["MergedAbbr"] <- NA#Insert a new column to the data set file with NA values. Below codes replaces accorning the NA values with our desired Merged site names
adb$MergedAbbr[as.factor(adb$Site.Name) == "Nankilabougou"] <- "NK"#Site Nakilabougou (NK) in Mali remains as NK
adb$MergedAbbr[as.factor(adb$Site.Name) == "vk5"] <- "VK"#VK5 and VK7 are merged as VK
adb$MergedAbbr[as.factor(adb$Site.Name) == "vk7"] <- "VK"#VK5 and VK7 are merged as VK
adb$MergedAbbr[as.factor(adb$Site.Name) == "somousso"] <- "SM"#Somousso in BF remains as SM
adb$MergedAbbr[as.factor(adb$Site.Name) == "Ahero"] <- "WK"#Ahero, Kakamega and Busia Teso are merged as WK
adb$MergedAbbr[as.factor(adb$Site.Name) == "Kakamega"] <- "WK"#Ahero, Kakamega and Busia Teso are merged as WK
adb$MergedAbbr[as.factor(adb$Site.Name) == "Busia/Teso"] <- "WK"#Ahero, Kakamega and Busia Teso are merged as WK
adb$MergedAbbr[as.factor(adb$Site.Name) == "KILIFI"] <- "CK"#Kwale, Kilifi and Malindi are merged as CK
```

Global and local distribution of *TEP1* genotypes and alleles in Africa

```
adb$MergedAbbr[as.factor(adb$Site.Name) == "KWALE"] <- "CK"#Kwale, Kilifi and Malindi are merged as CK
adb$MergedAbbr[as.factor(adb$Site.Name) == "MALINDI"] <- "CK"#Kwale, Kilifi and Malindi are merged as CK
adb$MergedAbbr[as.factor(adb$Site.Name) == "Nkolbisson"] <- "CM"#All CM sites are merged as CM
adb$MergedAbbr[as.factor(adb$Site.Name) == "Nkolkoumou"] <- "CM"#All CM sites are merged as CM
adb$MergedAbbr[as.factor(adb$Site.Name) == "Nkolondom"] <- "CM"#All CM sites are merged as CM
adb$MergedAbbr[as.factor(adb$Site.Name) == "Mfou"] <- "CM"#All CM sites are merged as CM
adb$MergedAbbr[as.factor(adb$Site.Name) == "Mvan"] <- "CM"#All CM sites are merged as CM
# I use genetic R package or subsetting (TEP1 genotype splitting) approach. Here, I use the TEP1 genotype splitting approach.
# First, I split up TEP1 genotype column into two other columns containing the respective alleles, and call the new data file adb.melt
adb.alleles <- colsplit(adb$Genotype, "/", c("Allele1", "Allele2"))
# Combine original and new data frame
adb <- cbind(adb, adb.alleles)
# In adb.melt file, melt variable columns Genotype, Allele1 and Allele2 together to form a new column called TEP1stutus, and another column called Allele
adb.melt <- melt(adb, measure.vars = c("Genotype", "Allele1", "Allele2"), variable.name="TEP1status", value.name= "Allele")
adb.melt<-subset (adb.melt, !TEP1status=="Genotype")#remove genotype class, because we only need to analyse Allele1 and Allele2 variables, which we need work with to calculate allele frequencies. This should be reproducible with genetic r package.
adb.melt <- subset(adb.melt, !SpeciesAbbr=="MS")#omit hybrids
AFsiteA<-ggplot(adb.melt, aes(x=factor(MergedAbbr,levels =c("NK","SM","VK","CM","WK","CK")),fill=factor(Allele)))
+geom_bar(position="fill", width=0.62)+facet_wrap(~Species)
#AFsiteEA
AFsiteA<- AFsiteA+ggtitle ("Allele frequencies in all ssp. in Africa") +
 xlab ("Country") + ylab("Allele frequency") + theme(legend.title = element_text(colour="black", size=10, face="bold")) +
 theme(legend.text = element_text(colour="black", size=10, face="italic"))
AFsiteA<-AFsiteA+ theme(legend.position="right")
#AFsiteA
AFsiteA<- AFsiteA+scale_fill_manual(values=c("#0072B2","#56B4E9", "#CC79A7","#009E73","#B2DF8A"),
                    name="Allele",
                    labels=c("R1","R2","R3", "S1","S2"))
AFsiteA#Fig. 2-11
#End#
```

Appendix 3. Equations in population genetics and R script used in this study

Equations and part of the R script that were used in population genetics

A Hardy Weinberg Equilibrium (HWE), Inbreeding and inbreeding coefficient

Assuming the Wrights Fisher model, an ideal panmictic diploid organisms, for example the mosquitoes, mate randomly such that haploid gametes fuse to form a zygote with equal chances to produce the progeny into the next population. Considering the example I used in the HWE introduction in chapter 1 of this thesis:

TEP locus can have two alleles, say R and S;
their allele frequencies will given by p and q respectively; and thus,
$q = 1- p$.
Genotype frequencies will be given by:
p^2 for the genotype of RR;
$2pq$ for the genotype of RS; and
q^2 for the genotype of RS.

Knowing the counts of individual genotypes that are observed in the population, the p and q allele frequencies can be calculated, and expected genotypes can be predicted by multiplying the individual observed genotype counts by their respective $p2$, $2pq$ and q^2 genotype frequencies.

In this context, inbreeding occurs when members of the same genotypes mate together, leading to higher probability (P) progeny of one observed genotype (G), say RR in this case, being identical by descent.

Inbreeding coefficient *(f)* measures relative deviation from the HWE expectations. The f is calculated as;

$$f = \frac{2pq - \mathrm{P}[RR = RS]}{2pq}$$

If the genotypes are not experiencing the inbreeding, but are instead mating preferentially (selectively) with other genotypes leading to high heterozygosities in the population, then the homozygote population is said to be breaking or undergoing the Wahlund effect.

B Fixation indices

Sewall Wright introduced three inbreeding coefficients that a structured population can have. These are called fixation indices or *F*-statistics namely: F_{IS}, F_{IT} and F_{ST}. F_{IS} measures deficiency of heterozygote individuals considering the allele frequencies in the subpopulation. F_{IT} measures deficiency of heterozygote individuals considering the allele frequencies in the total population. Therefore, both the F_{IS} and F_{IT} consider observed heterozygote individuals. The F_{ST} measures the difference between the expected and observed heterozygosity of subpopulations in comparison to the total population. The F_{ST} thus, takes into account allele frequency between demes (subpopulations), and may depend on species. Because it measures excess of homozygotes, its little value e.g. 0.2 in mosquito population may be considered huge and suggest that the population is structured.

Generally;

$f_x = \frac{H_{exp} - H_{obs}}{H_{exp}}$, where f_x can be F_{IS}, or F_{IT} or F_{ST}.

In this case, you can have two subpopulations, A and B.
Will be given by
$H_S = 2pq$ (Heterozygotes in A only), $H_T = \underline{2pq}$ (Sum of heterozygotes of A and B), and $H_I = \underline{p}$ (Observed frequency of heterozygotes of A).

Equations and part of the R script that were used in population genetics

It follows that;

$F_{IS} = \frac{H_S - H_I}{H_S}$, $F_{IT} = \frac{H_T - H_I}{H_T}$, $F_{ST} = \frac{H_T - H_S}{H_T}$ or $\frac{F_{IT} - F_{IS}}{1 - F_{IS}}$.

Below is a part of the R script that was used to calculate the inbreeding coefficient and the *F*-statistics, using Mali's two demes for sympatric *A. coluzzii* and *A. gambiae s.s.* populations.

C A part of an R script that I used to calculate the Fixation indices

```
# load packages
rm(list = ls())#clear environment
my_packages2 <- c("plyr", " genetics", " reshape2")
libraries(my_packages2)
getwd()#check the current wdir, and set it if necessary
# data set
db <- read.csv ("mvb.1.csv",sep=",",dec=".")#The column names in the dataset are Individual.ID","Country","Sample.ID","Site.Name","SpeciesAbbr","Species","Genotype".
#Abbreviation of countries by inserting a new column "CountryAbbr". The Country column has four levels: Mali, Burkina Faso, Cameroon and Kenya
db$CountryAbbr[as.factor(db$Country) == "Mali"] <- "ML"
db$CountryAbbr[as.factor(db$Country) == "Burkina Faso"] <- "BF"
db$CountryAbbr[as.factor(db$Country) == "Cameroon"] <- "CM"
db$CountryAbbr[as.factor(db$Country) == "Kenya"] <- "KE"
#Calculation of Mali local F statistics on the A.coluzzii population
ml<-subset(db, CountryAbbr =="ML")#Extract ML data only
mlc<-subset(ml, SpeciesAbbr =="M")#Extract A. coluzzii only
#Define Genotypes as objects using r-genetic package
mlc$Genotype <- genotype(mlc$Genotype)
#Step 1. Calculate allele frequencies using genetic package etc compare that you get these the figures. Compare that you get same figures as in Steps 6 and 7.
summary(mlc$Genotype)#check mosquito counts, allele and genotype frequencies; Genotype = c("R1/R1", "R1/S1", "R1/S2", "S1/S1", "S1/S2","S2/S2"), observed genotype frequencies= c(87,21,1,7,0,0)
R1f1<-(2*87+21+1)/(2*gc1)#[1] 0.8448276#Allele frequency of R1
S1f1<-(21+2*7+0)/(2*gc1)#[1] 0.1508621#Allele frequency of S1
S2f1<-(1+0+0)/(2*gc1)#[1] 0.004310345#Allele frequency of S1
R1f1+S1f1+S2f1#[1] 0.8448276#Global gene frequency is CORRECT if it adds up to 1
#Define those allele frequencies (p,q,r...) and total genotype counts (gc),
#mannual calculation of allele frequencies to four significant numbers may be better than using r genetic package (two significant numbers). Be sure to cross check accurately.
p1=0.8448#R1 allele frequency in 4 significant figures
q1=0.1509#S1 allele frequency in 4 significant figures
r1=0.0043#S2 allele frequency in 4 significant figures
gc1=116#Total number of genotyped mosquitoes
#Step 2 Calculate expected genotypic counts according to the Hardy Weinberg Equilibrium
c(gc1*p1^2,gc1*2*p1*q1,gc1*2*p1*r1,gc1*q1^2,gc1*2*q1*r1,gc1*r1^2)#[1]
82.78769664 29.57543424  0.84277248  2.64141396  0.15053784  0.00214484
#Create a data frame (MLc) containing the observed and the expected genotype frequencies
MLc <- data.frame(Country = "ML", Site ="NK", Population = "A. coluzzii",Genotype=
```

Equations and part of the R script that were used in population genetics

```
c("R1/R1", "R1/S1", "R1/S2", "S1/S1", "S1/S2","S2/S2"),
Observed= c(87,21,1,7,0,0), Expected
=c(gc1*p1^2,gc1*2*p1*q1,gc1*2*p1*r1,gc1*q1^2,gc1*2*q1*r1,gc1*r1^2))
#Calculate the HWE excess or deficiency of each genotype between the observed and the
expected frequencies
MLc["ObsExpDifference"]<-(MLc$Observed)-(MLc$Expected)
#Calculate homozygote excess or deficiency in relation to HWE
MLchm<-subset(MLc, !Genotype== "R1/S1")#Omit "R1/S1"
MLchm<-subset(MLchm, !Genotype== "R1/S2")#Omit "R1/S2"
MLchm<-subset(MLchm, !Genotype== "S1/S2")#Omit "S1/S2"
sum(MLchm$ObsExpDifference)
#[1] 8.568745
#Percentage Homozygote excess or deficiency# excess means its inbred and
#deficiency means its outbred or under Wahlund effect therefore the isolate is breaking
sum(MLchm$ObsExpDifference)/sum(MLchm$Expected)*100#Percentage Homozygote
excess
#[1] 10.02999 shows that there is excess of homozygotes hence inbreeding occurs
#Step 3 Calculate the local observed heterozgosities, Hobs. The genotypes are Counted
MLcht<-subset(MLc, !Genotype== "R1/R1")#Exclude "R1/R1"
MLcht<-subset(MLcht, !Genotype== "S1/S1")#Exclude "S1/S1"
MLcht<-subset(MLcht, !Genotype== "S2/S2")#Exclude "S2/S2"
HobsMLc<-sum(MLcht$Observed)/gc1#frequency of local observed heterozgosities
HobsMLc#[1] 0.1896552
#Step 4 Calculate local expected heterozygosity or gene diversity
HexpMLc<-1-(p1^2+q1^2+r1^2)#Frequency of local expected heterozygosity method2
HexpMLc#[1] 0.2635237 ie 26%
#Step 5 Calculate local (or global) inbreeding coefficient, Fs using step 3 and step 4
FsMLc<-(HexpMLc-HobsMLc)/HexpMLc
FsMLc#[1] 0.2803106 i.e. 28%
#Zero value means that the observed genotypes are according to HWE expectations
#positive Fs (fewer heterozygotes than expected)indicates inbreeding#
#negative Fs (more heteroyzgote than expected) shows excess out breeding
#Step 6 and 7 Calculate p-bar the frequency of R1 over the total population
R1f1<-(2*87+21+1)/(2*gc1)#[1] 0.8448276#Allele frequency of R1
S1f1<-(21+2*7+0)/(2*gc1)#[1] 0.1508621#Allele frequency of S1
S2f1<-(1+0+0)/(2*gc1)#[1] 0.004310345#Allele frequency of S1
R1f1+S1f1+S2f1#[1] 0.8448276#Global gene frequency is CORRECT if it adds up to 1
#Step 8 calculate global hererozygosity indices
#Calculation of Mali local F statistics on the A. gambiae s.s. subpopulation (The second
species in sympatry with the A. coluzzii)
mlg<-subset(ml, SpeciesAbbr =="S")#Extract A. gambiae s.s. subset
mlg$Genotype<-genotype(mlg$Genotype)#define "Genotype" column to be genotypes of
A. gambiae s.s. mosquitoes using the r-genetic package
summary(mlg$Genotype)#Check frequencies of mosquitoes, alleles and genotypes
#Step 1. Calculate allele frequencies using r genetic package etc and
#Define those allele frequencies(p,q,r...) and total genotypeCounts (gc)
#Define #Genotype = c("R1/R1", "R1/S1", "R1/R2","R2/R2", "R2/S1", "S1/S1")
#Define observed genotype frequencies#Observed= c(0,2,0,4,54,0,90,0,0)
gc2=150#total mosquitoes genotyped for this species
```

Equations and part of the R script that were used in population genetics

```
#calculation of allele frequencies. Cross check accurately.
p2=R1f<-(1*2)/(2*gc2)#[1] 0.006666667#Allele frequency of R1
q2=R2f<-((2*4)+(1*54))/(2*gc2)#[1] 0.2066667#Allele frequency of R2
r2=S1f<-((2*90)+(1*54)+(1*2))/(2*gc2)#[1] 0.7866667#Allele frequency of S1
p2+q2+r2#=1 Global gene frequency is CORRECT when it adds up to 1
#Step 2 #Calculate expected genotypic counts
c(gc2*p2^2,gc2*2*p2*q2,gc2*2*p2*r2,gc2*q2^2,gc2*2*q2*r2,gc2*r2^2)#[1]
0.006666667  0.413333333  1.573333333  6.406666667 48.773333333 92.826666667
#Create a data frame containing the observed and the expected genotype frequencies
MLg <- data.frame(Country = "ML", Site ="NK", Population = "A. gambiae s.s.",
Genotype= c("R1/R1", "R1/R2", "R1/S1","R2/R2", "R2/S1", "S1/S1"),
Observed= c(0,0,2,4,54,90), Expected
=c(gc2*p2^2,gc2*2*p2*q2,gc2*2*p2*r2,gc2*q2^2,gc2*2*q2*r2,gc2*r2^2))
#Calculate the HWE excess or deficiency of each genotype between the observed and the
expected frequencies
MLg["ObsExpDifference"]<-(MLg$Observed)-(MLg$Expected)
#Calculate homozygote excess or defieciency in relation to HWE
MLghm2<-subset(MLg, !Genotype == "R1/R2")#Exclude "R1/R2"
MLghm2<-subset(MLghm2, !Genotype == "R1/S1")#Exclude "R1/S1"
MLghm2<-subset(MLghm2, !Genotype == "R2/S1")#Exclude "R2/S1"
sum(MLghm2$ObsExpDifference)#[1] -5.24
#Percentage Hom excess or defiecieny# excess means its inbred and
#defeiciency means its outbred or under Wahlund effect therefore the isolate is breaking
sum(MLghm2$ObsExpDifference)/sum(MLghm2$Expected)*100#[1] -
5.280129#Percentage Homozygote excess
#Step 3 Calculate the local observed heterozgosities, Hobs. Count the genotypes
MLght2<-subset(MLg, !Genotype == "R1/R1")#Omit "R1/R1"
MLght2<-subset(MLght2, !Genotype == "R2/R2")#Omit "R2/R2"
MLght2<-subset(MLght2, !Genotype == "S1/S1")#Omit "S1/S1"
HobsMLg<-sum(MLght2$Observed)/gc2#frequency of local observed heterozgosities
HobsMLg#[1] 0.3733333
#Step 4 Calculate local expected heterozygosity or gene diversity
HexpMLg<- 1-(p2^2+q2^2+r2^2)#Frequency of local expected heterozygosity method2
HexpMLg#[1] 0.3384
#Step 5 Calculate local (or global) inbreeding coefficient, Fs using step 3 and step 4
Fs2MLg<-(HexpMLg-HobsMLg)/HexpMLg
Fs2MLg#[1] -0.1032309
#Zero means its according to HWE expectations
#positive Fs (fewer heterozygotes than expected)indicates inbreeding#
#negative Fs (more heteroyzgote than expected) shows excess out breeding
#Step 6 and 7 Calculate p-bar the frequency of R1 over the total population
R1f2<-(2*0+2+0)/(2*gc2)#[1] 0.006666667#Allele frequency of R1
R2f2<-(2*4+54+0)/(2*gc2)#[1] 0.2066667#Allele frequency of R2
S1f2<-(2+54+(2*90)+0)/(2*gc2)#[1] 0.7866667#Allele frequency of S1
R1f2+R2f2+S1f2#[1] 1
#Global gene frequency is CORRECT add upto 1
#Step 8 calculate global hererozygosity indices( over individuals, subpopulations and
Total population)
#First two calculations employ a weighted average of the values in the whole set of
```

Equations and part of the R script that were used in population genetics

subpopulations
#H1 based on observed heterozygosities in individuals in subpopulations
H1MLcg<- (HobsMLc*116+HobsMLg*150)/(116+150)
H1MLcg#[1] 0.2932331
#Hs based on expected heterozygosities in subpopulations
HsMLcg<- (HexpMLc*116+HexpMLg*150)/(116+150)
HsMLcg#[1] 0.3384
#Ht based on expected heterozygosities for overall total populations-using global allele frequencies
MLR1bar = ((p1*232)+(p2*300))/532#[1] 0.3721684
MLR2bar = ((q2*300))/532#[1] 0.1165414
MLS1bar = ((q1*232)+(r2*300))/532#[1] 0.509415
MLS2bar = ((r1*232))/532#[1] 0.001875188
MLR1bar+MLR2bar+MLS1bar+MLS2bar#[1] 1
HtMLcg<- 1-((MLR1bar)^2+(MLR2bar)^2+(MLS1bar)^2+(MLS2bar)^2)
HtMLcg#[1] 0.5884016
#Step 9 Calculate global F-statistics
#Compare and contrast the global Fis below with the 'local inbreeding coefficient' Fs of step 5.
#Here we are using a weighted average of the individual heterozygosities over all the subpopulations.
#Both Fis and Fs are based on the observed heterozygosities,
#where as Fst and Fit are based on the expected heterozygosities
FisMLcg <- (HsMLcg-H1MLcg)/HsMLcg#[1] 0.133472#Fis
FstMLcg <- (HtMLcg-HsMLcg)/HtMLcg#[1] 0.4248826#Fst
FitMLcg <- (HtMLcg-H1MLcg)/HtMLcg#[1] 0.5016446#Fit
#End#

Appendix 4. Statistical Tests for the Hardy Weinberg Equilibrium

Statistical Tests for the Hardy Weinberg Equilibrium

Chi-square tests on the HWE. Asterisk (*) indicates significant deviation from the HWE at $\chi 2_{0.05,}\ p>0.05$.
$\chi 2_{\text{Cal}}$ (1) = standard $\chi 2$, while $\chi 2_{\text{Cal}}$ (2) = conservative $\chi 2$ that corrects for small sample size. See pages 22-23.

	Country	Site	Species	Genotype	Obs	Exp	Obs-Exp	$\chi 2_{\text{Cal}}$ (1)	$\chi 2_{\text{Cal}}$ (2)	$\chi 2_{0.05}$
1	ML	NK	*A. col*	*R1/R1*	87	82.788	4.212	0.214	0.17	
2	ML	NK	*A. col*	*R1/S1*	21	29.575	-8.575	2.486	2.2	
3	ML	NK	*A. col*	*R1/S2*	1	0.843	0.157	0.029	0.14	
4	ML	NK	*A. col*	*S1/S1*	7	2.641	4.359	7.192	5.64	
5	ML	NK	*A. col*	*S1/S2*	0	0.151	-0.151	0.151	0.81	
6	ML	NK	*A. col*	*S2/S2*	0	0.002	-0.002	0.002	115.56	
							df = 3	∑10.07*	∑124.52*	7.815
7	ML	NK	*A. gam*	*R1/R1*	0	0.007	-0.007	0.007	36.51	
8	ML	NK	*A. gam*	*R1/R2*	0	0.413	-0.413	0.413	0.02	
9	ML	NK	*A. gam*	*R1/S1*	2	1.573	0.427	0.116	0	
10	ML	NK	*A. gam*	*R2/R2*	4	6.407	-2.407	0.904	0.57	
11	ML	NK	*A. gam*	*R2/S1*	54	48.773	5.227	0.560	0.46	
12	ML	NK	*A. gam*	*S1/S1*	90	92.827	-2.827	0.086	0.06	
							df = 3	∑2.086	∑37.61*	7.815
13	BF	SM	*A. col*	*R2/R2*	2	2.083	-0.083	0.003	0.08	
14	BF	SM	*A. col*	*R2/S1*	6	5.833	0.167	0.005	0.02	
15	BF	SM	*A. col*	*S1/S1*	4	4.083	-0.083	0.002	0.04	
							df = 1	∑0.010	∑0.14	3.841
16	BF	SM	*A. gam*	*R1/R1*	0	0.005	-0.005	0.005	46.01	
17	BF	SM	*A. gam*	*R1/R2*	0	0.351	-0.351	0.351	0.06	
18	BF	SM	*A. gam*	*R1/S1*	1	0.638	0.362	0.205	0.03	
19	BF	SM	*A. gam*	*R2/R2*	9	5.793	3.207	1.776	1.27	
20	BF	SM	*A. gam*	*R2/S1*	15	21.064	-6.064	1.746	1.47	
21	BF	SM	*A. gam*	*S1/S1*	22	19.149	2.851	0.424	0.29	
							df = 3	∑4.508	∑49.12*	7.815
22	CM	MV	*A. col*	*R2/R2*	0	0.316	-0.316	0.316	0.11	
23	CM	MV	*A. col*	*R2/S1*	6	6.456	-0.456	0.032	0	
24	CM	MV	*A. col*	*R2/S2*	4	2.911	1.089	0.407	0.12	
25	CM	MV	*A. col*	*S1/S1*	34	32.924	1.076	0.035	0.01	
26	CM	MV	*A. col*	*S1/S2*	28	29.696	-1.696	0.097	0.05	
27	CM	MV	*A. col*	*S2/S2*	7	6.696	0.304	0.014	0.01	
							df = 3	∑0.901	∑0.29	7.815
28	CM	MV	*A. gam*	*R2/R2*	0	0.593	-0.593	0.593	0.01	
29	CM	MV	*A. gam*	*R2/S1*	8	5.630	2.370	0.998	0.62	
30	CM	MV	*A. gam*	*R2/S2*	0	1.185	-1.185	1.185	0.4	
31	CM	MV	*A. gam*	*S1/S1*	13	13,370	-0.370	0.010	0	
32	CM	MV	*A. gam*	*S1/S2*	4	1.653	2.347	3.331	2.06	
33	CM	MV	*A. gam*	*S2/S2*	2	0.593	1.407	3.343	1.39	
							df = 3	∑9.459*	∑4.49	7.815
34	CM	NS	*A. col*	*R2/R2*	2	1.359	0.641	0.,303	0.01	
35	CM	NS	*A. col*	*R2/S1*	16	16.413	-0.413	0.010	0	
36	CM	NS	*A. col*	*R2/S2*	5	5.870	-0.870	0.129	0.02	
37	CM	NS	*A. col*	*S1/S1*	49	49.567	-0.567	0.006	0	
38	CM	NS	*A. col*	*S1/S2*	37	35.452	1.548	0,068	0.03	
39	CM	NS	*A. col*	*S2/S2*	6	6.339	-0.339	0.018	0	
							df = 3	∑0.534	∑0.07	7.815
40	CM	NS	*A. gam*	*R2/R2*	6	8.975	-2.975	0.986	0.68	
41	CM	NS	*A. gam*	*R2/S1*	86	76.698	9.302	1.128	1.01	
42	CM	NS	*A. gam*	*R2/S2*	1	4.352	-3.352	2.581	1.87	
43	CM	NS	*A. gam*	*S1/S1*	159	163.85	-4.854	0.144	0.12	
44	CM	NS	*A. gam*	*S1/S2*	19	4.960	14.040	39.74	36.96	
45	CM	NS	*A. gam*	*S2/S2*	2	0.527	1.473	4.111	1.79	
							df = 3	∑48.69*	∑42.43*	7.815

Statistical Tests for the Hardy Weinberg Equilibrium

Chi-square tests on the HWE. Asterisk (*) indicates significant deviation from the HWE at $\chi 2_{0.05,}\ p>0.05$.
$\chi 2$ Cal (1) = standard $\chi 2$, while $\chi 2$ Cal (2) = conservative $\chi 2$ that corrects for small sample size. See pages 22-23.

	Country	Site	Species	Genotype	Obs	Exp	Obs-Exp	$\chi 2$ Cal (1)	$\chi 2$ Cal (2)	$\chi 2_{0.05}$
46	CM	MF+ ND+ NM	*A. col*	*R2/R2*	0	0.766	-0.766	0.766	0.09	
47	CM	MF+ ND+ NM	*A. col*	*R2/S1*	6	5.031	0.969	0.187	0.04	
48	CM	MF+ ND+ NM	*A. col*	*R2/S2*	1	0,438	0,563	0.723	0.01	
49	CM	MF+ ND+ NM	*A. col*	*S1/S1*	8	8.266	-0.266	0.009	0.01	
50	CM	MF+ ND+ NM	*A. col*	*S1/S2*	1	1.438	-0.438	0.133	0	
51	CM	MF+ ND+ NM	*A. col*	*S2/S2*	0	0.063	-0.063	0.063	3.06	
								df = 3	∑1.880	∑3.22	7.815
52	CM	MF+ ND+ NM	*A. gam*	*R2/R2*	6	11.207	-5.207	2.419	1.98	
53	CM	MF+ ND+ NM	*A. gam*	*R2/S1*	106	97.067	8.933	0.822	0.73	
54	CM	MF+ ND+ NM	*A. gam*	*R2/S2*	6	4.519	1.481	0.485	0.21	
55	CM	MF+ ND+ NM	*A. gam*	*S1/S1*	207	210.181	-3.181	0.048	0.03	
56	CM	MF+ ND+ NM	*A. gam*	*S1/S2*	17	5.167	11.833	27.102	24.86	
57	CM	MF+ ND+ NM	*A. gam*	*S2/S2*	1	0.456	0.544	0.651	0	
								df = 3	∑31.52*	∑27.82*	7.815
58	KE	AH	*A. ara*	*R2/R2*	19	9.600	9.400	9.204	8.25	
59	KE	AH	*A. ara*	*R2/S1*	10	28.800	-18.800	12.27	11.63	
60	KE	AH	*A. ara*	*S1/S1*	31	21.600	9.400	4.091	3.67	
								df = 1	∑25.56*	∑23.55*	3.841
61	KE	AH	*A. gam*	*R2/R2*	6	3.521	2.479	1.746	1.11	
62	KE	AH	*A. gam*	*R2/S1*	1	5.958	-4.958	4.126	3.34	
63	KE	AH	*A. gam*	*S1/S1*	5	2.521	2.479	2.438	1.55	
								df = 1	∑8.310*	∑6*	7.815
64	KE	BT	*A. gam*	*R2/R2*	4	2.196	1.804	1.482	0.77	
65	KE	BT	*A. gam*	*R2/S1*	18	22.446	-4.446	0.881	0.69	
66	KE	BT	*A. gam*	*R2/S2*	1	0.163	0.837	4.311	0.7	
67	KE	BT	*A. gam*	*S1/S1*	60	57.361	2.639	0.121	0.08	
68	KE	BT	*A. gam*	*S1/S2*	0	0.831	-0.831	0.831	0.13	
69	KE	BT	*A. gam*	*S2/S2*	0	0.003	-0.003	0.003	82	
								df = 3	∑7.630	∑84.38*	7.815
70	KE	KK	*A. ara*	*R2/R2*	0	0.800	-0.800	0.800	0.11	
71	KE	KK	*A. ara*	*R2/S1*	1	2.400	-1.400	0.817	0.34	

Statistical Tests for the Hardy Weinberg Equilibrium

Chi-square tests on the HWE. Asterisk (*) indicates significant deviation from the HWE at $\chi 2_{0.05,}\ p>0.05$.
$\chi 2_{\text{Cal}}$ (1) = standard $\chi 2$, while $\chi 2_{\text{Cal}}$ (2) = conservative $\chi 2$ that corrects for small sample size. See pages 22-23.

	Country	Site	Species	Genotype	Obs	Exp	Obs-Exp	$\chi 2_{\text{Cal}}$ (1)	$\chi 2_{\text{Cal}}$ (2)	$\chi 2_{0.05}$
72	KE	KK	*A. ara*	*S1/S1*	4	1.800	2.200	2.689	1.61	
							df = 1	∑4.306*	∑2.06	3.841
73	KE	KK	*A. gam*	*R2/R2*	2	0.333	1.667	8.333	4.08	
74	KE	KK	*A. gam*	*R2/S1*	2	5.333	-3.333	2.083	1.51	
75	KE	KK	*A. gam*	*S1/S1*	23	21.333	1.667	0.130	0.06	
							df = 1	∑10,54*	∑5.65*	3.841
76	KE	MD	*A. ara*	*R2/R2*	4	3.571	0.429	0.051	0	
77	KE	MD	*A. ara*	*R2/S1*	2	2.,857	-0.857	0.257	0.04	
78	KE	MD	*A. ara*	*S1/S1*	1	0.571	0.429	0.321	0.01	
							df = 1	∑0.630	∑0.06	3.841
79	KE	MD	*A. mer*	*R2/R2*	3	1.455	1.545	1.642	0.75	
80	KE	MD	*A. mer*	*R2/R3*	0	2.000	-2.000	2.000	1.13	
81	KE	MD	*A. mer*	*R2/S1*	10	10.727	-0.727	0.049	0	
82	KE	MD	*A. mer*	*R2/S2*	0	0.364	-0.364	0.364	0.05	
83	KE	MD	*A. mer*	*R3/R3*	3	0.688	2.313	7.778	4.78	
84	KE	MD	*A. mer*	*R3/S1*	5	7.375	-2.375	0.765	0.48	
85	KE	MD	*A. mer*	*R3/S2*	0	0.250	-0.250	0.250	0.25	
86	KE	MD	*A. mer*	*S1/S1*	22	19.778	2.222	0.250	0.15	
87	KE	MD	*A. mer*	*S1/S2*	0	1.341	-1.341	1.341	0.53	
88	KE	MD	*A. mer*	*S2/S2*	1	0.023	0.977	42.02	10.02	
							df = 6	∑56.46*	∑18.14*	12.59

Appendix 5. *TEP*R3* full-length nucleotide alignment with other allele sequences

```
*R1_ML       1 ATGTGGCAGTTCATAAGGTCACGAATATTAACGGTGATAATCTTCATAGGTGCTGCTCAT
*R1_L3_5     1 ............................................................
*R2_4Arr     1 ............................................................
*R2_ML       1 ............................................................
*R3_KE       1 ............................................................
*R3_KE.      1 ............................................................
*S1_G3S3     1 ............................................................
*S1_CM       1 ............................................................
*S2_4Arr     1 ............................................................
*S2_CM       1 ............................................................
               Intron 1
*R1_ML      61 GGGTAGGAACAAACGTGCTGGAAGTGCTGTGCCAATCGATTGAGTTGAGAGTAATTTGTA
*R1_L3_5    61 .------------------------------------------------------------
*R2_4Arr    61 ............................................................
*R2_ML      61 ............................................................
*R3_KE      61 ............................................................
*R3_KE.     61 .....................T......................................
*S1_G3S3    61 .------------------------------------------------------------
*S1_CM      61 ..........................................A.............A..
*S2_4Arr    61 ..............................----..........................
*S2_CM      61 ..........................................A.................

*R1_ML     121 CACTACGAAACGAAATATACTTTTTCTAGGCTACTGGTTGTGGGTCCGAAATTTATACGG
*R1_L3_5    62 --------------------------.............................
*R2_4Arr   121 ..........T.....................................................C...
*R2_ML     121 ..........................................................C...
*R3_KE     121 ..........................................................C...
*R3_KE.    121 ..........................................................C...
*S1_G3S3    62 --------------------------.............................C...
*S1_CM     121 ..........................................................C...
*S2_4Arr   117 ..........................................................C...
*S2_CM     121 ..........................................................C...

*R1_ML     181 GCCAACCAGGAATACACTCTGGTGATCAGCAACTTTAACTCACAGCTAAGCAAAGTGGAC
*R1_L3_5    94 ............................................................
*R2_4Arr   181 ............................................................
*R2_ML     181 ............................................................
*R3_KE     181 ............................................................
*R3_KE.    181 ............................................................
*S1_G3S3    94 ............................................................
*S1_CM     181 ............................................................
*S2_4Arr   177 ............................................................
*S2_CM     181 ............................................................

*R1_ML     241 CTGCTGTTAAAACTGGAAGGCGAAACTGATAATGGTTTAAGCGTTCTGAACGTTACCAAG
*R1_L3_5   154 ............................................................
*R2_4Arr   241 ............................................................
*R2_ML     241 ............................................................
*R3_KE     241 ............................................................
*R3_KE.    241 ............................................................
*S1_G3S3   154 ...T........................................................
*S1_CM     241 ...T........................................................
*S2_4Arr   237 ............................................................
*S2_CM     241 ...T........................................................
                                                          Intron 2
*R1_ML     301 ATGGTTGACGTGCGACGTAATATGAACCGAATGATCAACTTCAATGTATGAAGAGTGAGC
*R1_L3_5   214 ..........................................-----------------
*R2_4Arr   301 ............................................................
*R2_ML     301 ............................................................
*R3_KE     301 ............................................................
*R3_KE.    301 ............................................................
*S1_G3S3   214 ..........................................-----------------
*S1_CM     301 ............................................................
*S2_4Arr   297 ............................................................
*S2_CM     301 ............................................................
```


The full-length sequences of *TEP1*R3* (highlighted in grey color) compared to **R1* and **R2* alleles from Mali, **R2* and **S1* alleles from Cameroon, and **R1,*R2, *S1* alleles from insectary strains. Intronic regions **(in bold)** with modifications for *TEP1*R3* are highlighted in Turquoise color.

```
*R1_ML     361 GATATTAGTTTTTAAGGCTTACAACTAAAACATTCGATCCTTTGCAGATGCCTGAGGATC
*R1_L3_5   259 -------------------------------------------.............
*R2_4Arr   361 ............................................................
*R2_ML     361 ............................................................
*R3_KE     361 A...........................................................
*R3_KE.    361 A...........................................................
*S1_G3S3   259 -------------------------------------------.............
*S1_CM     361 ....................T.......................................
*S2_4Arr   357 ..........C.................................................
*S2_CM     361 ............................................................

*R1_ML     421 TGACGGCTGGAAACTACAAAATAACTATCGATGGACAGCGTGGCTTCAGCTTTCACAAGG
*R1_L3_5   272 ............................................................
*R2_4Arr   421 ............................................................
*R2_ML     421 ............................................................
*R3_KE     421 ........................................A...................
*R3_KE.    421 ........................................A...................
*S1_G3S3   272 ............................................................
*S1_CM     421 ............................................................
*S2_4Arr   417 ............................................................
*S2_CM     421 ............................................................

*R1_ML     481 AGGCAGAGCTGGTGTATCTCAGCAAATCGATATCGGGGCTAATACAGGTCGATAAGCCCG
*R1_L3_5   332 ............................................................
*R2_4Arr   481 ............................................................
*R2_ML     481 ............................................................
*R3_KE     481 ............................................................
*R3_KE.    481 ............................................................
*S1_G3S3   332 ............................................................
*S1_CM     481 ............................................................
*S2_4Arr   477 ............................................................
*S2_CM     481 ............................................................

*R1_ML     541 TATTTAAACCTGGGGATACGGTGAACTTCCGTGTGATCGTGCTGGACACGGAGCTGAAAC
*R1_L3_5   392 ............................................................
*R2_4Arr   541 ............................................................
*R2_ML     541 ............................................................
*R3_KE     541 ............C...........................................G.
*R3_KE.    541 ............C...........................................G.
*S1_G3S3   392 ............................................................G.
*S1_CM     541 ............................................................G.
*S2_4Arr   537 ............................................................G.
*S2_CM     541 ............................................................G.

*R1_ML     601 CGCCGGCGAGGGTCAAGTCGGTTTATGTAACTATACGAGATCCTCAGCGCAATGTGATTC
*R1_L3_5   452 ............................................................
*R2_4Arr   601 ............................................................
*R2_ML     601 ............................................................
*R3_KE     601 .A..A.......................................................
*R3_KE.    601 ......................C.....................................
*S1_G3S3   452 ......................C.....................................
*S1_CM     601 ......................C.....................................
*S2_4Arr   597 ....A.......................................................
*S2_CM     601 ......................C.....................................

*R1_ML     661 GCAAATGGTCCACGGCAAAACTGTATGCCGGTGTGTTCGAGAGCGATCTACAGATAGCGC
*R1_L3_5   512 ............................................................
*R2_4Arr   661 ..............................................G.............
*R2_ML     661 ..............................................G.............
*R3_KE     661 ..............................................G........A....
*R3_KE.    661 ..............................................G........A....
*S1_G3S3   512 ..............................................G.............
*S1_CM     661 ..............................................G.............
*S2_4Arr   657 ..................G...........................G.............
*S2_CM     661 ..............................................G.............
```


The full-length sequences of *TEP1*R3* (highlighted in grey color) compared to **R1* and **R2* alleles from Mali, **R2* and **S1* alleles from Cameroon, and **R1,*R2, *S1* alleles from insectary strains. Intronic regions **(in bold)** with modifications for *TEP1*R3* are highlighted in Turquoise color.

```
*R1_ML     721 CTACTCCAATGCTCGGGGTCTGGAATATCTCGGTGGAGGTGGAAGGAGAAGAGCTTGTGT
*R1_L3_5   572 ............................................................
*R2_4Arr   721 ..........................................G.................
*R2_ML     721 ..........................................G.................
*R3_KE     721 ..........................................G.................
*R3_KE.    721 ............................................................
*S1_G3S3   572 ..............................A.............................
*S1_CM     721 ..............................A.............................
*S2_4Arr   717 ............................................................
*S2_CM     721 ..............................A.............................

*R1_ML     781 CAAAGACGTTTGAGGTGAAGGAGTACGTGTTGTCAACGTTCGACGTGCAGGTCATGCCAT
*R1_L3_5   632 ............................................................
*R2_4Arr   781 ........................T...................................
*R2_ML     781 ........................T...................................
*R3_KE     781 ............................................................
*R3_KE.    781 ............................................................
*S1_G3S3   632 ................A...........................................
*S1_CM     781 ................A...........................................
*S2_4Arr   777 ............................................................
*S2_CM     781 ................A...........................................

*R1_ML     841 CGGTGATTCCACTGGAAGAGCATCAAGCTGTGAATCTTACAATCGAAGCGAACTATCACT
*R1_L3_5   692 ............................................................
*R2_4Arr   841 ......................C.......................C.............
*R2_ML     841 ......................C.......................C.............
*R3_KE     841 ......................C..........C........................
*R3_KE.    841 ......................C..........C........................
*S1_G3S3   692 ............................................................
*S1_CM     841 ............................................................
*S2_4Arr   837 ............................................................
*S2_CM     841 ............................................................

*R1_ML     901 TTGGTAAGCCAGTGCAAGGAGTGGCCAAGGTGGAGCTGTACCTAGACGACGATAAGCTAA
*R1_L3_5   752 ............................................................
*R2_4Arr   901 ............................................................
*R2_ML     901 ............................................................
*R3_KE     901 ........................T.....................A.............
*R3_KE.    901 ........................T.....................A.............
*S1_G3S3   752 ............................................................
*S1_CM     901 ............................................................
*S2_4Arr   897 ............................................................
*S2_CM     901 ............................................................

*R1_ML     961 AACTGAAAAAAGAGCTGACTGTGTACGGAAAGGGCCAGGTAGAGTTGCGCTTTGACAATT
*R1_L3_5   812 ............................................................
*R2_4Arr   961 .T.A..................A.......................A.....C.......
*R2_ML     961 .T.A..................A.......................A.....C.......
*R3_KE     961 .T.A.................................................A......
*R3_KE.    961 .T.A.................................................A......
*S1_G3S3   812 .T.AA........................................................
*S1_CM     961 .T.AA........................................................
*S2_4Arr   957 .T.A.........................................................
*S2_CM     961 .T.AA........................................................

*R1_ML    1021 TTGCAATGGATGCGGATCAGCAGGATGTACCAGTGAAGGTGTCGTTCATCGAGCAGTACA
*R1_L3_5   872 ..............................................G.............
*R2_4Arr  1021 ............................G.G.........................C...
*R2_ML    1021 ............................G.G.........................C...
*R3_KE    1021 ............A..............G.............A..T.........C...
*R3_KE.   1021 ............................G.............A..T.........C...
*S1_G3S3   872 ............................G.............................
*S1_CM    1021 ............................G.............................
*S2_4Arr  1017 ............................G.............................
*S2_CM    1021 ............................G.............................
```


The full-length sequences of *TEP1*R3* (highlighted in grey color) compared to **R1* and **R2* alleles from Mali, **R2* and **S1* alleles from Cameroon, and **R1,*R2, *S1* alleles from insectary strains. Intronic regions **(in bold)** with modifications for *TEP1*R3* are highlighted in Turquoise color.

```
                  Intron 3
*R1_ML    1081 CAAGTAAGAATCATGTTCGAGATACCGTTGCTAACAGTGATTAAATAAGAAATATCTCTT
*R1_L3_5   932 ...----------------------------------------------------------
*R2_4Arr  1081 .....................A....................................
*R2_ML    1081 .....................A....................................
*R3_KE    1081 .......C.............................................
*R3_KE.   1081 .......C.............................................
*S1_G3S3   932 ...----------------------------------------------------------
*S1_CM    1081 .....................A...........T........................
*S2_4Arr  1077 .....................A...........T........................
*S2_CM    1081 .....................A...........T........................

*R1_ML    1141 CATAGATCGTACGGTGGTCAAACAGTCACAAATCACGGTATATAGGTATGCGTACCGAGT
*R1_L3_5   935 -----.....................................................
*R2_4Arr  1141 .....................................................
*R2_ML    1141 .....................................................
*R3_KE    1141 ..C..................................................
*R3_KE.   1141 ..C..................................................
*S1_G3S3   935 -----.....................................................
*S1_CM    1141 .....................................................
*S2_4Arr  1137 .....................................................
*S2_CM    1141 .....................................................

*R1_ML    1201 AGAGTTGATAAAAGAGAGTCCACAGTTTCGTCCGGGACTCCCGTTCAAATGTGCGCTTCA
*R1_L3_5   990 .....................................................
*R2_4Arr  1201 .....................................................G..
*R2_ML    1201 .....................................................G..
*R3_KE    1201 .....................................................G..
*R3_KE.   1201 .....................................................G..
*S1_G3S3   990 .....................................................G..
*S1_CM    1201 .....................................................G..
*S2_4Arr  1197 ...A.................................................A..G..
*S2_CM    1201 .....................................................G..

*R1_ML    1261 GTTTACACACCATGATGGAACACCGGCTAAAGGCATTAGCGGTAAGGTAGAGGTATCCGA
*R1_L3_5  1050 .....................................................
*R2_4Arr  1261 ..................................C...G...............
*R2_ML    1261 ..................................C...G...............
*R3_KE    1261 .....................................................
*R3_KE.   1261 .....................................................
*S1_G3S3  1050 ..................................C...G............T..
*S1_CM    1261 ..................................C...G............T..
*S2_4Arr  1257 ..................................C...G...............
*S2_CM    1261 ..................................C...G...............

*R1_ML    1321 TGTACGATTCGAAACGACAACAACGAGTGATAACGATGGATTGATTAAGCTCGAGCTGCA
*R1_L3_5  1110 .....................................................
*R2_4Arr  1321 ....G................................................
*R2_ML    1321 ....G................................................
*R3_KE    1321 ....G................................................
*R3_KE.   1321 ....G................................................
*S1_G3S3  1110 ....G..............A.....................................
*S1_CM    1321 ....G................................................
*S2_4Arr  1317 ....G................................................
*S2_CM    1321 ....G................................................
                                                  Intron 4
*R1_ML    1381 ACCAAGTGAGGGTACTGAACAACTCAGTATTCACGTAAGTATCTAGAATGTTTAGTTAAT
*R1_L3_5  1170 ..............................----------------------------
*R2_4Arr  1381 ........................G.....A..................G...T.....
*R2_ML    1381 ........................G.....A..................G...T.....
*R3_KE    1381 ........................G.....AG.........G......G...T.....
*R3_KE.   1381 ........................G.....AG.........G......G...T.....
*S1_G3S3  1170 ............T...........G.....A..----------------------------
*S1_CM    1381 ........................G.....A..................G...T.....
*S2_4Arr  1377 ........................G.....A..................G...T.....
*S2_CM    1381 ........................G.....A..................G...T.....
```


The full-length sequences of *TEP1*R3* (highlighted in grey color) compared to **R1* and **R2* alleles from Mali, **R2* and **S1* alleles from Cameroon, and **R1,*R2, *S1* alleles from insectary strains. Intronic regions **(in bold)** with modifications for *TEP1*R3* are highlighted in Turquoise color.

```
*R1_ML    1441 GGTTGACAAAGCATCTTAAAGGGTCAGTTCTTTTGCAGTTCAATGCTGTTGATGGATTCT
*R1_L3_5  1204 -----------------------------------.....................
*R2_4Arr  1441 ................G.....................................
*R2_ML    1441 ................G.....................................
*R3_KE    1441 C...............G.....................................
*R3_KE.   1441 C...............G.....................................
*S1_G3S3  1204 -----------------------------------.....................
*S1_CM    1441 ................G.....................................
*S2_4Arr  1437 ................G.....................................
*S2_CM    1441 ................G.....................................

*R1_ML    1501 TTTTTTATGAAGATGTGAATAACGTAGAAACGGTTACAAATGCGTATATTAAACTGGAGC
*R1_L3_5  1226 ....................G.............GG.....................
*R2_4Arr  1501 ....................G.............GG.....................
*R2_ML    1501 ....................G.............GG.....................
*R3_KE    1501 ....................G.............GG....T..C.............
*R3_KE.   1501 ....................G.............GG....T..C.............
*S1_G3S3  1226 ....................G.............GG.....................
*S1_CM    1501 ....................G.............GG.....................
*S2_4Arr  1497 ....................G.............GG.....................
*S2_CM    1501 ....................G.............GG.....................
                        Intron 5
*R1_ML    1561 TGAAATCACCGTGAGTAATAACTCGCTACAAAGTGAAACTGGCAGTGTGATGTATAACAT
*R1_L3_5  1286 .........-------------------------------------------------
*R2_4Arr  1561 ..........................................................
*R2_ML    1561 ..........................................................
*R3_KE    1561 ....................T.............GG.....................
*R3_KE.   1561 ....................T.............GG.....................
*S1_G3S3  1286 .........-------------------------------------------------
*S1_CM    1561 ..................................................A.......
*S2_4Arr  1557 ..................................................A.......
*S2_CM    1561 ..................................................A.......

*R1_ML    1621 AACATATCGTTCTAGCATCAAACGGAACAAATTGATGCGTTTCATGGTGACGTGCACGGA
*R1_L3_5  1296 --------------.....................................C.....
*R2_4Arr  1621 ....................................C.....................
*R2_ML    1621 ....................................C.....................
*R3_KE    1621 ....................................C.....................
*R3_KE.   1621 ....................................C.....................
*S1_G3S3  1296 --------------.....................C.....................
*S1_CM    1621 ....................................C.....................
*S2_4Arr  1617 ....................................C.....................
*S2_CM    1621 ....................................C.....................

*R1_ML    1681 GCGCATGACATTCTTCGTGTACTATGTCATGTCAAAGGGCAATATCATCGATGCAGGCTT
*R1_L3_5  1341 ..........................................................A..
*R2_4Arr  1681 ..........................................................
*R2_ML    1681 ..........................................................
*R3_KE    1681 ..............T.........................C..T..............
*R3_KE.   1681 ..............T.........................C..T..............
*S1_G3S3  1341 ..........................................................
*S1_CM    1681 ..........................................................
*S2_4Arr  1677 ..........................................................
*S2_CM    1681 ..........................................................

*R1_ML    1741 CATGCGACCCAACAAGCAACCGAAGTACCTGTTGCAGCTGAACGCAACAGAAAAGATGAT
*R1_L3_5  1401 ..........................................................
*R2_4Arr  1741 ..................A.......T...........A...................
*R2_ML    1741 ..................A...................A...................
*R3_KE    1741 ..................A...................A...................
*R3_KE.   1741 ..................A...................A...................
*S1_G3S3  1401 ..................A...................A...................
*S1_CM    1741 ..................A...................A...................
*S2_4Arr  1737 ..................A...................A...................
*S2_CM    1741 ..................A...................A...................
```


The full-length sequences of *TEP1*R3* (highlighted in grey color) compared to **R1* and **R2* alleles from Mali, **R2* and **S1* alleles from Cameroon, and **R1,*R2, *S1* alleles from insectary strains. Intronic regions **(in bold)** with modifications for *TEP1*R3* are highlighted in Turquoise color.

```
*R1_ML    1801 TCCGAGGGCGAAAATTCTCATCGCTACCGTAGCGGGCCGCACGGTGGTGTACGACTTCGC
*R1_L3_5  1461 ............................................................
*R2_4Arr  1801 .....A......................................................
*R2_ML    1801 .....A......................................................
*R3_KE    1801 .....A......................................................
*R3_KE.   1801 .....T......................................................
*S1_G3S3  1461 .....A.....................................A..........A...
*S1_CM    1801 .....A.....................................A..........A...
*S2_4Arr  1797 .....A.....................................A..........A...
*S2_CM    1801 .....A.....................................A..........A...

                                                 Intron 6
*R1_ML    1861 AGACCTCGATTTCCAAGAGCTTCGCAATAATGTAAGCATTTGTTTGTCGTGTTGTTTAAC
*R1_L3_5  1521 ............................-----------------------------
*R2_4Arr  1861 ............................................................
*R2_ML    1861 ............................................................
*R3_KE    1861 ............................................................
*R3_KE.   1861 ....................................A....A....AC.........G
*S1_G3S3  1521 ............................-----------------------------
*S1_CM    1861 ............................................................
*S2_4Arr  1857 ............................................................
*S2_CM    1861 ............................................................

*R1_ML    1921 GTAACACTTATTCATGTTGTGTGGAAACAGTTTGATTTAAGCATTGACGAGCAAGAGATC
*R1_L3_5  1552 ------------------------------............................
*R2_4Arr  1921 ....A.......................................................
*R2_ML    1921 ....A.......................................................
*R3_KE    1921 ....A.......................................................
*R3_KE.   1921 ............G......T..A.....................................
*S1_G3S3  1552 ------------------------------............................
*S1_CM    1921 ............................................................
*S2_4Arr  1917 ............................................................
*S2_CM    1921 ............................................................

*R1_ML    1981 AAGCCGGGACGACAAATCGAGCTGAGCATGTCTGGACGCCCAGGAGCGTACGTTGGGCTG
*R1_L3_5  1582 ............................................................
*R2_4Arr  1981 ..............................T.............................
*R2_ML    1981 ............................................................
*R3_KE    1981 ............................................................
*R3_KE.   1981 ............................................................
*S1_G3S3  1582 ............................................................
*S1_CM    1981 ............................................................
*S2_4Arr  1977 ............................................................
*S2_CM    1981 ............................................................

*R1_ML    2041 GCCGCGTATGACAAAGCCTTGCTGCTTTTCAACAAGAACCACGACCTGTTCTGGGAGGAC
*R1_L3_5  1642 ............................................................
*R2_4Arr  2041 ........................G...................................
*R2_ML    2041 ........................G...................................
*R3_KE    2041 ......................A..G...................................
*R3_KE.   2041 ......................A..G...................................
*S1_G3S3  1642 ........................G...................................
*S1_CM    2041 ........................G...................................
*S2_4Arr  2037 ........................G...................................
*S2_CM    2041 ........................G...................................

*R1_ML    2101 ATTGGGCAGGTGTTTGATGGGTTCCATGCAATCAATGAGAACGAGTTTGACATATTCCAC
*R1_L3_5  1702 ............................................................
*R2_4Arr  2101 ............................................................
*R2_ML    2101 ............................................................
*R3_KE    2101 ............................................................
*R3_KE.   2101 ............................................................
*S1_G3S3  1702 ............................................................
*S1_CM    2101 ............................................................
*S2_4Arr  2097 ............................................................
*S2_CM    2101 ............................................................
```


The full-length sequences of *TEP1*R3* (highlighted in grey color) compared to **R1* and **R2* alleles from Mali, **R2* and **S1* alleles from Cameroon, and **R1,*R2, *S1* alleles from insectary strains. Intronic regions **(in bold)** with modifications for *TEP1*R3* are highlighted in Turquoise color.

```
                Intron 7
*R1_ML    2161 GTATGTATGATGCGAAAATCGAGCAAGAG------ATATCAGAA----------------
*R1_L3_5  1762 ------------------------------------------------------------
*R2_4Arr  2161 ..........................------.........----------------
*R2_ML    2161 ..........................------.........----------------
*R3_KE    2161 .............T.............CAAGAT...C.....TTTATCACAAACAATT
*R3_KE.   2161 .............A.............CAAGAT...C.....TTTATCACAAACAATT
*S1_G3S3  1762 ------------------------------------------------------------
*S1_CM    2161 ..........................------.........----------------
*S2_4Arr  2157 ..........................------.........----------------
*S2_CM    2161 ..........................------.........----------------

*R1_ML    2199 AACAATTATCAAAAACGAGACGCGTAATTATTTTTGCAGAGCTTGGGTCTGTTCGCCAGG
*R1_L3_5  1762 ------------------------------------.....................
*R2_4Arr  2199 ........................G....C.............................
*R2_ML    2199 ........................G....C.............................
*R3_KE    2221 .T.........C.....G............C.............................
*R3_KE.   2221 .T.........C.....G............C.............................
*S1_G3S3  1762 ------------------------------------.....................
*S1_CM    2199 ............................C.............................
*S2_4Arr  2195 ............................C.............................
*S2_CM    2199 ............................C.............................

*R1_ML    2259 ACATTGGACGATATCTTGTTCGACAGTGCAAATGAAAAGACGGGGCGTAATGCACTGCAG
*R1_L3_5  1783 ............................................................
*R2_4Arr  2259 ............................................................
*R2_ML    2259 ............................................................
*R3_KE    2281 ...C..................T.....................................
*R3_KE.   2281 ...C........................................................
*S1_G3S3  1783 ............................................................
*S1_CM    2259 ............................................................
*S2_4Arr  2255 ............................................................
*S2_CM    2259 ............................................................

*R1_ML    2319 TCAGGCAAGCCGATCGGCAAGCTGGTGTCGTATCGGACGAACTTCCAGGAATCGTGGTTG
*R1_L3_5  1843 ............................................................
*R2_4Arr  2319 ............................................................
*R2_ML    2319 ............................................................
*R3_KE    2341 ............................................................
*R3_KE.   2341 ............................................................
*S1_G3S3  1843 ............................................................
*S1_CM    2319 ............................................................
*S2_4Arr  2315 ............................................................
*S2_CM    2319 ............................................................

*R1_ML    2379 TGGAAAAATGTTTCCATCGGACGATCGGGAAGTCGCAAGTTGATCGAGGTAGTACCGGAC
*R1_L3_5  1903 ............................................................
*R2_4Arr  2379 ............................................................
*R2_ML    2379 ............................................................
*R3_KE    2401 ............................................................
*R3_KE.   2401 ............................................................
*S1_G3S3  1903 ............................................................
*S1_CM    2379 ............................................................
*S2_4Arr  2375 ............................................................
*S2_CM    2379 ............................................................

*R1_ML    2439 ACGACCACCTCCTGGTATCTGACGGGCTTCTCGATCGATCCCGTGTACGGGTTGGGTATC
*R1_L3_5  1963 ............................................................
*R2_4Arr  2439 ..A.....T.............C.....................................
*R2_ML    2439 ..A.....T.............C.....................................
*R3_KE    2461 ............................................................
*R3_KE.   2461 ............................................................
*S1_G3S3  1963 ............................................................
*S1_CM    2439 ............................................................
*S2_4Arr  2435 ............................................................
*S2_CM    2439 ............................................................
```


The full-length sequences of *TEP1*R3* (highlighted in grey color) compared to **R1* and **R2* alleles from Mali, **R2* and **S1* alleles from Cameroon, and **R1,*R2, *S1* alleles from insectary strains. Intronic regions **(in bold)** with modifications for *TEP1*R3* are highlighted in Turquoise color.

```
*R1_ML    2499 ATCAAGAAGCCAATCCAGTTCACAACAGTCCAGCCGTTCTACATCGTAGAGAACTTACCA
*R1_L3_5  2023 ............................................................
*R2_4Arr  2499 ............................................................
*R2_ML    2499 ............................................................
*R3_KE    2521 ............................................................
*R3_KE.   2521 ............................................................
*S1_G3S3  2023 ............................................................C.....
*S1_CM    2499 ............................................................C.....
*S2_4Arr  2495 ............................................................TC.....
*S2_CM    2499 ............................................................TC.....

*R1_ML    2559 TATTCAATCAAACGAGGCGAAGCGGTTGTGTTGCAGTTTACGCTGTTCAACAACCTTGGA
*R1_L3_5  2083 ............................................................
*R2_4Arr  2559 ............................................................
*R2_ML    2559 ............................................................
*R3_KE    2581 ........................C...................................
*R3_KE.   2581 ........................C...................................
*S1_G3S3  2083 ........................C...................................
*S1_CM    2559 ........................C...................................
*S2_4Arr  2555 ........................C...................................
*S2_CM    2559 ........................C...................................

*R1_ML    2619 GCGGAGTATATAGCCGATGTGACGCTGTACAATGTGGCCAACCAGACCGAGTTCGTCGGA
*R1_L3_5  2143 ............................................................
*R2_4Arr  2619 ............................................................
*R2_ML    2619 ............................................................
*R3_KE    2641 ............................................................
*R3_KE.   2641 ............................................................
*S1_G3S3  2143 ............................................................
*S1_CM    2619 ............................................................
*S2_4Arr  2615 ............................................................
*S2_CM    2619 ............................................................

                                  Intron 8
*R1_ML    2679 CGTCCAAATACGGGTGAGTGTGGTTTACATCAATCAACCCTTGATTATT-GAAAACTTCA
*R1_L3_5  2203 ............----------------------------------------------
*R2_4Arr  2679 ............................................................
*R2_ML    2679 ............................................................
*R3_KE    2701 ......GT............A.....GT......A.------.....CATA....AA.A.
*R3_KE.   2701 ......GT............A.....GT.........GA.....C...A-.....AA.A.
*S1_G3S3  2203 ......G......----------------------------------------------
*S1_CM    2679 ......G............AA.....GT......A.------.....C--AT...AAAA.
*S2_4Arr  2675 ......G............AA.....GT......A.------.....C--AT...AAAA.
*S2_CM    2679 ......G............AA.....GT......A.------.....C--AT...AAA-.

*R1_ML    2738 ACATTAATTTTATGTTCAGATCTCAGCTACACCAAATCCGTGAGCGTTCCTCCAAAAGTT
*R1_L3_5  2216 ------------------..........................................
*R2_4Arr  2738 ............................................................
*R2_ML    2738 ............................................................
*R3_KE    2755 .....T.C..----.C............G.........G..A...........G..G...
*R3_KE.   2760 .....T....----.C............G.........G..A...........G..G...
*S1_G3S3  2216 ------------------..........T.........G..A...........G..G...
*S1_CM    2731 .....T....----.C............T.........G..A...........G..G...
*S2_4Arr  2727 .....T....----.C............T.........G..T...........G..G...
*S2_CM    2730 .....T....----.C............T.........G..T...........G..G...

*R1_ML    2798 GGTGTGCCAATCTCGTTCCTCATCAAGGCCCGCAAGCTCGGCGAGATGGCGGTTCGTGTA
*R1_L3_5  2257 ............................................................
*R2_4Arr  2798 ............................................................
*R2_ML    2798 ............................................................
*R3_KE    2811 .....T..G.....A..TT..G.A.................................G
*R3_KE.   2816 .....T..G.....A..TT..G.A.................................G
*S1_G3S3  2257 .....T..G.....T..T.....A.............................A.....G
*S1_CM    2787 .....T..G.....T..T.....A.............................A.....G
*S2_4Arr  2783 .....T..G.....T..T.....A.............................A.....G
*S2_CM    2786 .....T..G.....T..T.....A.............................A.....G
```


The full-length sequences of *TEP1*R3* (highlighted in grey color) compared to **R1* and **R2* alleles from Mali, **R2* and **S1* alleles from Cameroon, and **R1,*R2, *S1* alleles from insectary strains. Intronic regions **(in bold)** with modifications for *TEP1*R3* are highlighted in Turquoise color.

```
*R1_ML    2858 AAGGCTTCGATAATGCTGGGACACGAAACGGACGCCCTGGAAAAGGTAATACGGGTGATG
*R1_L3_5  2317 ............................................................
*R2_4Arr  2858 ............................................................
*R2_ML    2858 ............................................................
*R3_KE    2871 ..................G...........T..GT.........G...........
*R3_KE.   2876 ..................G...........T..GT.........G...........
*S1_G3S3  2317 ..........................A.....AT.........G...A........
*S1_CM    2847 ..........................A.....AT.........G...A........
*S2_4Arr  2843 ..........................A..T..AT.........G...A........
*S2_CM    2846 ..........................A..T..AT.........G...A........

*R1_ML    2918 CCTGAAAGTTTGGTGCAGCCGAGAATGGATACACGCTTTTTCTGCTTCGACGATTACAAA
*R1_L3_5  2377 ...........................................................C.....
*R2_4Arr  2918 ............................................................
*R2_ML    2918 ............................................................
*R3_KE    2931 ..C.....C...........................................................
*R3_KE.   2936 ..C.....C...........................................................
*S1_G3S3  2377 ..C..........C...A..A.A..........A.....................
*S1_CM    2907 ..C..........C...A..A.A..........A.....................
*S2_4Arr  2903 ..C..........C...A..A.A..........A.....................
*S2_CM    2906 ..C..........C...A..A.A..........A.....................

*R1_ML    2978 AATCAAACGTTTTCGATCAACTTGGACATCAACAAGAAGGCCGACAGTGGATCGACAAAG
*R1_L3_5  2437 ............C...............................................
*R2_4Arr  2978 ............C.................................A.............
*R2_ML    2978 ............C.................................A.............
*R3_KE    2991 ............C.......T.........................A........A....
*R3_KE.   2996 ............C.......T.........................A........A....
*S1_G3S3  2437 ............CC.TT.............................T.A........A....
*S1_CM    2967 ............CC.TT.............................T.A........A....
*S2_4Arr  2963 ............CC.TT.............................T.A........A....
*S2_CM    2966 ............CC.TT.............................T.A........A....

                                  Intron 9
*R1_ML    3038 ATTGAGTTTCGACTAAATCGTAAGTAGAGGGTGT-GAAAGTTGTGAAAGGAGTTATTGAG
*R1_L3_5  2497 ..................----------------------------------------
*R2_4Arr  3038 ...................................-......................
*R2_ML    3038 ...................................-......................
*R3_KE    3051 ....................................A.....A...G....C..A.....
*R3_KE.   3056 ....................................A.....A...G....C..A.....
*S1_G3S3  2497 .....A...A.......C.----------------------------------------
*S1_CM    3027 .....A...A.......C.......TA.A.....AC...C.A............T.....
*S2_4Arr  3023 .....A...A.......C.......TA.A.....AC...C.A............T.....
*S2_CM    3026 .....A...A.......C.......TA.A.....AC...C.A............T.....

*R1_ML    3097 AGTTTTTTTCTTCTTTTTACCCAACCATTTCTGCAGCCAATTTGTTGACCACGGTCATCA
*R1_L3_5  2516 ----------------------------------.......................
*R2_4Arr  3097 ............................................................
*R2_ML    3097 ............................................................
*R3_KE    3111 .T......AAA.TC..........T..............C...C......T........
*R3_KE.   3116 .T......AAA.TC..........T..............C...C......T........
*S1_G3S3  2516 ------------------------------------....C...C......T........
*S1_CM    3087 ........A--.TC.A.....TG.T..............C...C......T........
*S2_4Arr  3083 ........A--.TC.A.....TG.T..............C...C......T........
*S2_CM    3086 ........A--.TC.A.....TG.T..............C...C......T........

*R1_ML    3157 AGAACCTGGACCATCTTCTCGGCGTTCCGACGGGATGTGGTGAGCAGAATATGGTCAAAT
*R1_L3_5  2540 ............................................................
*R2_4Arr  3157 ............................................................
*R2_ML    3157 ............................................................
*R3_KE    3171 .......A...A....A....C.........................A..C.........
*R3_KE.   3176 .......A...A....A....C.........................A..C.........
*S1_G3S3  2540 ...........A....A....C.....................................
*S1_CM    3145 ...........A....A....C.....................................
*S2_4Arr  3141 ...........A....A....C.....................................
*S2_CM    3144 ...........A....A....C.....................................
```


The full-length sequences of *TEP1*R3* (highlighted in grey color) compared to **R1* and **R2* alleles from Mali, **R2* and **S1* alleles from Cameroon, and **R1,*R2, *S1* alleles from insectary strains. Intronic regions **(in bold)** with modifications for *TEP1*R3* are highlighted in Turquoise color.

```
*R1_ML    3217 TTGTTCCCAACATTTTGGTGCTGGATTATTTGCATGCCATCGGGTCGAAAGAACAGCATC
*R1_L3_5  2600 ..................A.....................................
*R2_4Arr  3217 ........................................................
*R2_ML    3217 ........................................................
*R3_KE    3231 ........................................................
*R3_KE.   3236 ........................................................
*S1_G3S3  2600 ..........................C..T......C...A.....G..........
*S1_CM    3205 ..........................C..T......C...A.....G..........
*S2_4Arr  3201 ..........................C..T......C...A.....G..........
*S2_CM    3204 ..........................C..T......C...A.....G..........

*R1_ML    3277 TAATCGACAAAGCTACGAATTTGTTGCGACAAGGATATCAAAACCAGATGCGCTACCGTC
*R1_L3_5  2660 ..........................T.............................
*R2_4Arr  3277 ..........................T.............................
*R2_ML    3277 ..........................T.............................
*R3_KE    3291 ..........................T.........................T....
*R3_KE.   3296 ..........................T.........................T....
*S1_G3S3  2660 ..........................T..G........G...........T.....C.
*S1_CM    3265 ..........................T..G........G...........T.....C.
*S2_4Arr  3261 ..........................T..G........G...........T.....C.
*S2_CM    3264 ..........................T..G........G...........T.....C.

*R1_ML    3337 AGACGGATGGTTCATTTGGTTTGTGGGAGACTACTAATGGTAGCGTGTTTCTCACCGCGT
*R1_L3_5  2720 ........................................................
*R2_4Arr  3337 ................................GG......................
*R2_ML    3337 ................................GG......................
*R3_KE    3351 ..............................T..GG.........A............
*R3_KE.   3356 ..............................T..GG.........A............
*S1_G3S3  2720 ....A........G......G.........AA.G.GGCA.C................
*S1_CM    3325 ....A........G......G.........AA.G.GGCA.C................
*S2_4Arr  3321 ....A........G......G.........AA.G.GGCA.C................
*S2_CM    3324 ....A........G......G.........AA.G.GGCA.C................

*R1_ML    3397 TCGTTGGCACATCGATGCAAACTGCAGTAAATTACATAAGCGATATTGATGCAGCAGTGG
*R1_L3_5  2780 ............................A.......................A...
*R2_4Arr  3397 ..........................C...A.....................A...
*R2_ML    3397 ..........................C...A.....................A...
*R3_KE    3411 ........................TTC...A.......A......A.........A...
*R3_KE.   3416 ........................TTC...A.......A......A.........A...
*S1_G3S3  2780 ......C..................TCG..A.....G.A..............A...
*S1_CM    3385 ......C..................TCG..A.....G.A..............A...
*S2_4Arr  3381 ......C..................TCG..A.....G.A..............A...
*S2_CM    3384 ......C..................TCG..A.....G.A..............A...

*R1_ML    3457 TGGAGAAGGCATTGGATTGGTTAGCCTCGAAGCAGCATTTCTCGGGACGGTTTGACAAGG
*R1_L3_5  2840 ........................................................
*R2_4Arr  3457 ........................................................
*R2_ML    3457 ........................................................
*R3_KE    3471 ................................A....C..................
*R3_KE.   3476 ................................A....C..................
*S1_G3S3  2840 .A.................G.............CAG......G.........G..A
*S1_CM    3445 .A.................G.............CAG......G.........G..A
*S2_4Arr  3441 .A.................G.............CAG......G.........G..A
*S2_CM    3444 .A.................G.............CAG......G.........G..A

*R1_ML    3517 CCGGTGCAGAGTATCACAAAGAAATGCAAGGAGGGTTGCGCAATGGTGTGGCCCTCACAT
*R1_L3_5  2900 ........................................................
*R2_4Arr  3517 ........................................................
*R2_ML    3517 ........................................................
*R3_KE    3531 ........................................................
*R3_KE.   3536 ........................................................
*S1_G3S3  2900 .....AA..T..GG........T..................................
*S1_CM    3505 .....AA..T..GG........T..................................
*S2_4Arr  3501 .....AA..T..GG........T..................................
*S2_CM    3504 .....AA..T..GG........T..................................
```


The full-length sequences of *TEP1*R3* (highlighted in grey color) compared to **R1* and **R2* alleles from Mali, **R2* and **S1* alleles from Cameroon, and **R1,*R2, *S1* alleles from insectary strains. Intronic regions **(in bold)** with modifications for *TEP1*R3* are highlighted in Turquoise color.

```
*R1_ML    3577 CATATGTGTTGATGGCATTGCTGGAGAATGACATCGCCAAAGCAAAGCACGCAGAGGTGA
*R1_L3_5  2960 ................................T.......................
*R2_4Arr  3577 ............................................................
*R2_ML    3577 ............................................................
*R3_KE    3591 ............C.........A...........T.......................
*R3_KE.   3596 ............C.........A...........T...................C.....
*S1_G3S3  2960 .G..........C..............................TG........G.T...A.
*S1_CM    3565 .G..........C..............................TG.......TG.T...A.
*S2_4Arr  3561 .G..........C..............................TG........G.T...A.
*S2_CM    3564 .G..........C..............................TG........G.T...A.

*R1_ML    3637 TTCAAAAAGGAATGACCTATCTGAGCAATCAGTTTGGATCCATCAACAATGCATACGACC
*R1_L3_5  3020 ............................................................
*R2_4Arr  3637 ............................................................
*R2_ML    3637 ............................................................
*R3_KE    3651 ............................................................
*R3_KE.   3656 .C.............A.............................G...........T....
*S1_G3S3  3020 .C.....C.......A.................C...C..T...T...............
*S1_CM    3625 .C.....C.......A.................C...C..T...T...............
*S2_4Arr  3621 .C.....C.......A...............AC...C..T...T...............
*S2_CM    3624 .C.....C.......A...............AC...C..T...T...............

*R1_ML    3697 TATCGATAGCAACCTACGCGATGATGTTGAACGGACACACCATGAAGGAGGAGGCACTCA
*R1_L3_5  3080 ............................................................
*R2_4Arr  3697 ............................................................
*R2_ML    3697 ............................................................
*R3_KE    3711 ............................................................G.
*R3_KE.   3716 ............................C...............................G.
*S1_G3S3  3080 ...............................................A.A.........G
*S1_CM    3685 ...............................................A.A.........G
*S2_4Arr  3681 ............................C..................A.A.........G
*S2_CM    3684 ............................C..................A.A.........G

*R1_ML    3757 ATAAGCTGATTGATATGTCTTTCATTGATGCTGATAAAAACGAACGGTTCTGGAACACAA
*R1_L3_5  3140 ............................................................
*R2_4Arr  3757 ............................................................
*R2_ML    3757 ............................................................
*R3_KE    3771 ....A.............................C..................T....GA....
*R3_KE.   3776 ....A.............................C..................T....GA....
*S1_G3S3  3140 ......................A...G....AA.A.......A.......A....GGA....
*S1_CM    3745 ......................A...G....AA.A.......A.......A....GGA....
*S2_4Arr  3741 ............................................................
*S2_CM    3744 ............................................................

*R1_ML    3817 CGAATCCAATAGAAACCACCGCATATGCTCTGCTGTCGTTTGTGATGGCCGAGAAGTACA
*R1_L3_5  3200 ............................................................
*R2_4Arr  3817 ............................................................
*R2_ML    3817 ............................................................
*R3_KE    3831 A.....A..................C................................TC
*R3_KE.   3836 A.....A..................C................................TC
*S1_G3S3  3200 ....C.A............A..............................A........TT
*S1_CM    3805 ....C.A............A..............................A........TT
*S2_4Arr  3801 ...............................................A........TT
*S2_CM    3804 ...............................................A........TT

*R1_ML    3877 CAGACGGTATACCGGTCATGAATTGGTTGGTGAATCAACGTTACGTTACCGGTAGCTTTC
*R1_L3_5  3260 ............................................................
*R2_4Arr  3877 ............................................................
*R2_ML    3877 ............................................................
*R3_KE    3891 TG..........................................................
*R3_KE.   3896 TG..........................................................
*S1_G3S3  3260 TG............A.T.......................G...........A.......
*S1_CM    3865 TG............A.T.......................G...........A.......
*S2_4Arr  3861 TG..............T.......................G...........A.......
*S2_CM    3864 TG..............T.......................G...........A.......
```


The full-length sequences of *TEP1*R3* (highlighted in grey color) compared to **R1* and **R2* alleles from Mali, **R2* and **S1* alleles from Cameroon, and **R1,*R2, *S1* alleles from insectary strains. Intronic regions **(in bold)** with modifications for *TEP1*R3* are highlighted in Turquoise color.

```
*R1_ML    3937 CGAGCACGCAAGACACGTTTGTGGGGCTGAAAGCGCTGACCAAAATGGCGGAAAAGATAT
*R1_L3_5  3320 ............................................................
*R2_4Arr  3937 ............................................................
*R2_ML    3937 ............................................................
*R3_KE    3951 ........................C...................................
*R3_KE.   3956 ........................C...................................
*S1_G3S3  3320 .AC......................C..T..............T...............
*S1_CM    3925 .AC......................C..T..............T...............
*S2_4Arr  3921 .AC......................C..T..............T...............
*S2_CM    3924 .AC......................C..T..............T...............

*R1_ML    3997 CTCCGTCCCGAAACGACTACACCGTTCAACTGAAGTACAAGAAGTGTGCAAAATACTTCA
*R1_L3_5  3380 ..........................................A.................
*R2_4Arr  3997 ..........................................A.................
*R2_ML    3997 ..........................................A.................
*R3_KE    4011 ..........................................A.................
*R3_KE.   4016 .......A................................A.................
*S1_G3S3  3380 ....................................T......A.CA....G.......
*S1_CM    3985 ....................................T......A.CA....G.......
*S2_4Arr  3981 ....................................T......A.CA....G.......
*S2_CM    3984 ....................................T......A.CA....G.......

*R1_ML    4057 AAATAAACTCGGAGCAAATTGATGTGGAAAACTTCGTGGGTATACCGGAGGACACAAAAA
*R1_L3_5  3440 ....................................A.......................
*R2_4Arr  4057 ............................................................
*R2_ML    4057 ............................................................
*R3_KE    4071 ....C.............................T...A....................
*R3_KE.   4076 ....C.............................T...A....................
*S1_G3S3  3440 .C..C..................T.CC....T...T...AA..............T....
*S1_CM    4045 .C..C..................T.CC....T...T...AA..............T....
*S2_4Arr  4041 .C..C.....................CC....T...T...AA..............G....
*S2_CM    4044 .C..C.....................CC....T...T...AA..............G....

*R1_ML    4117 AGCTCGAGATCAATGTGGGGGGCATTGGATTTGGGTTGTTAGAGGTGGTTTATCAATTTG
*R1_L3_5  3500 ............................................................A
*R2_4Arr  4117 ............................................................A
*R2_ML    4117 ............................................................A
*R3_KE    4131 ..........................................G.................C.
*R3_KE.   4136 ..........................................G................AC.
*S1_G3S3  3500 ....T..........A.....T.....G.....A...C.G.....CA............
*S1_CM    4105 ....T..........A.....T.....G.....A...C.G.....CA............
*S2_4Arr  4101 ....T..........A..A..T.....G.....A...C.G.....CA............
*S2_CM    4104 ....T..........A..A..T.....G.....A...C.G.....CA............

*R1_ML    4177 ATTTGAATCTCGTCAACTTTGAGAATAGATTCCAACTAGACCTGGAGAAACAGAACACAG
*R1_L3_5  3560 ............................................................
*R2_4Arr  4177 ............................................................
*R2_ML    4177 ............................................................
*R3_KE    4191 ............................................................
*R3_KE.   4196 ............................................................
*S1_G3S3  3560 ....A..................C..C.....A.........................
*S1_CM    4165 ....A..................C..C.....A.........................
*S2_4Arr  4161 ....A..................C..C.....A.........................
*S2_CM    4164 ....A..................C..C.....A.........................

*R1_ML    4237 GCTCTGACTACGAGCTGAGGCTGAAGGTCTGTGCCAGCTACATACCCCAGCTGACCGACA
*R1_L3_5  3620 ............................................................
*R2_4Arr  4237 ............................................................
*R2_ML    4237 ............................................................
*R3_KE    4251 ..................A....G..T.............C...A.............
*R3_KE.   4256 ..................A....G..T.............C...A.............
*S1_G3S3  3620 ..................A....G...........A......C...G...........
*S1_CM    4225 ..................A....G...........A......C...G...........
*S2_4Arr  4221 ..................A....G...........A......C...G...........
*S2_CM    4224 ..................A....G...........A......C...G...........
```


The full-length sequences of *TEP1*R3* (highlighted in grey color) compared to **R1* and **R2* alleles from Mali, **R2* and **S1* alleles from Cameroon, and **R1,*R2, *S1* alleles from insectary strains. Intronic regions **(in bold)** with modifications for *TEP1*R3* are highlighted in Turquoise color.

```
*R1_ML    4297 GACGATCGAACATGGCACTGATTGAGGTAACCTTACCGAGCGGTTACGTGGTTGATCGCA
*R1_L3_5  3680 ............................................................
*R2_4Arr  4297 ............................................................
*R2_ML    4297 ............................................................
*R3_KE    4311 ......................C.....G.....................T..C.......
*R3_KE.   4316 ......................C.....G.....................T..C.......
*S1_G3S3  3680 .T.A................A..C.....G.......................C.......
*S1_CM    4285 .T.A................A..C.....G.......................C.......
*S2_4Arr  4281 .T.A................A..C.....G.......................C.......
*S2_CM    4284 .T.A................A..C.....G.......................C.......
                                                  Intron 10
*R1_ML    4357 ATCCGATCAGCGAGCAGACGAAGGTGAATCCGATTCAGGTAAGAATATTTGAATGTTGAA
*R1_L3_5  3740 ..................................-------------------------
*R2_4Arr  4357 ............................................................
*R2_ML    4357 ............................................................
*R3_KE    4371 ............................................................
*R3_KE.   4376 ..................A.C........TA............................
*S1_G3S3  3740 ....A..............C............-------------------------
*S1_CM    4345 ....A..............C............A........C...........C..T.
*S2_4Arr  4341 ....A..............C............A........C...........C..T.
*S2_CM    4344 ....A..............C............A........C...........C..T.

*R1_ML    4417 TATCCAGAGCAGTTTGAGCTGACTATTATGTATTTACTTTTGATTGCATTCACAGAAAAC
*R1_L3_5  3775 ------------------------------------------------.........
*R2_4Arr  4417 ............................................................
*R2_ML    4417 ............................................................
*R3_KE    4431 ............................................................
*R3_KE.   4436 .......GA..A.....................................................
*S1_G3S3  3774 ------------------------------------------------......C.T
*S1_CM    4405 A..GG..GAT.C..G...GC----.CC..T..A..G...GGA...T.....G.....C.T
*S2_4Arr  4401 A..GG..GAT.C..G...GC----.CC..T..A..G...GGA...TT....G.....C.T
*S2_CM    4404 A..GG..GAT.C..G...GC----.CC..T..A..G...GGA...TT....G.....C.T

*R1_ML    4477 TGAAATCCGTTACGGTGGCACTTCAGTCGTTTTATACTACGACAATATGGGCAGCGAGCG
*R1_L3_5  3783 ............................................................
*R2_4Arr  4477 .........C..................................................
*R2_ML    4477 .........C..................................................
*R3_KE    4491 ............................................................
*R3_KE.   4496 ..............................C.G.....A.....................
*S1_G3S3  3783 G.....T.....................C.G......T............C......
*S1_CM    4461 G.....T.....................C.G......T............C......
*S2_4Arr  4457 G.....T.....................C.G......T............C......
*S2_CM    4460 G.....T.....................C.G......T............C......

*R1_ML    4537 TAACTGTTTCACCCTGACCGCGTACAGACGCTTTAAGGTCGCATTGAAGCGTCCAGCGTA
*R1_L3_5  3843 ............................................................
*R2_4Arr  4537 ............................................................
*R2_ML    4537 ............................................................
*R3_KE    4551 ............................................................
*R3_KE.   4556 ............AA.........................A.....................
*S1_G3S3  3843 .........T...G....T.....................................
*S1_CM    4521 A........T...G....T.....................................
*S2_4Arr  4517 A........T...G....T..................A..................
*S2_CM    4520 A........T...G....T..................A..................
                                              Intron 11
*R1_ML    4597 TGTGGTTGTGTATGATTATTATAATACAAGTGAGTAGTAGTCATAGATTGGCTATGGAAT
*R1_L3_5  3903 ..........................------------------------------
*R2_4Arr  4597 ............................................................
*R2_ML    4597 ............................................................
*R3_KE    4611 ............................................................
*R3_KE.   4616 ...T........C..C......G.....................................
*S1_G3S3  3903 ...T........C..............------------------------------
*S1_CM    4581 ...T........C.....................................A.........
*S2_4Arr  4577 ...T........C.............................A.....A.........
*S2_CM    4580 ...T........C.............................A.....A.........
```


The full-length sequences of *TEP1*R3* (highlighted in grey color) compared to **R1* and **R2* alleles from Mali, **R2* and **S1* alleles from Cameroon, and **R1,*R2, *S1* alleles from insectary strains. Intronic regions **(in bold)** with modifications for *TEP1*R3* are highlighted in Turquoise color.

```
*R1_ML    4657 TGCACAGGGAATG----------TAACACCCGTTGCTT--TTTAAATTATTTACAGATCT
*R1_L3_5  3932 ------------------------------------------------------....
*R2_4Arr  4657 ............----------............--..................
*R2_ML    4657 ............----------............--..................
*R3_KE    4671 ..........T..GTTATGGAATAG............--..................
*R3_KE.   4676 .A........T..GTTATGGAAT.G....T......A.--..................
*S1_G3S3  3932 ------------------------------------------------------..T.
*S1_CM    4641 .........C...----------.G....T........T-..................T.
*S2_4Arr  4637 .........T...----------.G....T.A......TT...T..............T.
*S2_CM    4640 .........T...----------.G....T.A......TT...T..............T.

*R1_ML    4705 GAACGCCATCAAAGTGTACGAAGTGGACAAGCAGAATTTGTGCGAAATCTGTGACGAAGA
*R1_L3_5  3936 ............................................................
*R2_4Arr  4705 ............................................................
*R2_ML    4705 ............................................................
*R3_KE    4729 ............................................................
*R3_KE.   4734 ............................................................
*S1_G3S3  3936 ....................................G.............C..G.....
*S1_CM    4690 ....................................G.............C..G.....
*S2_4Arr  4687 ....................................G.............C..G.....
*S2_CM    4690 ....................................G.............C..G.....

*R1_ML    4765 AGACTGTCCTGCAGAGTGC
*R1_L3_5  3996 ...................
*R2_4Arr  4765 ...................
*R2_ML    4765 ...................
*R3_KE    4789 ...................
*R3_KE.   4794 ...................
*S1_G3S3  3996 ...................
*S1_CM    4750 .........C.........
*S2_4Arr  4747 ...................
*S2_CM    4750 .........C.........
```


The full-length sequences of *TEP1*R3* (highlighted in grey color) compared to **R1* and **R2* alleles from Mali, **R2* and **S1* alleles from Cameroon, and **R1,*R2, *S1* alleles from insectary strains. Intronic regions **(in bold)** with modifications for *TEP1*R3* are highlighted in Turquoise color.

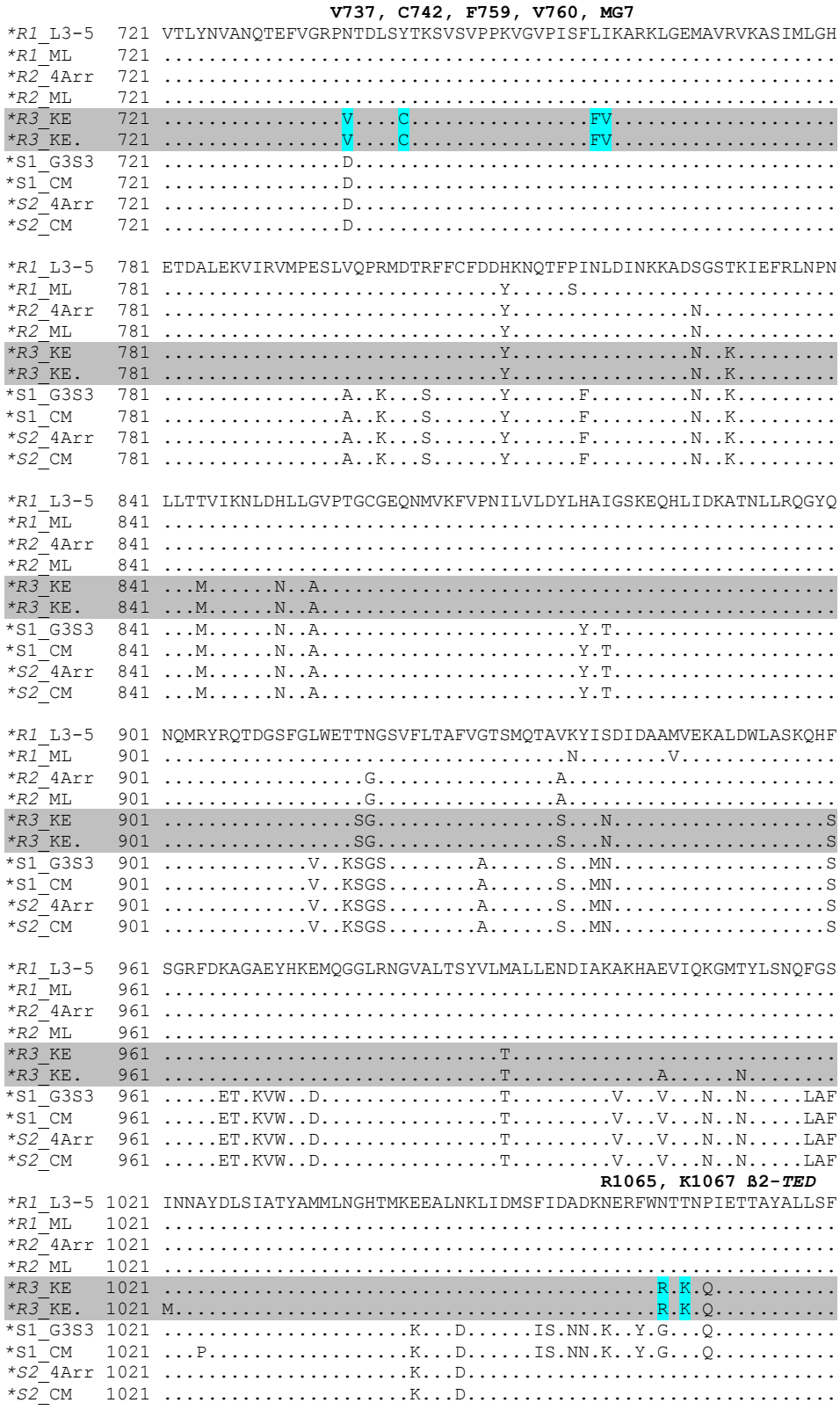
Appendix 6. *TEP*R3* full-length amino acid sequence alignment with other alleles

```
*R1_L3-5      1 MWQFIRSRILTVIIFIGAAHGLLVVGPKFIRANQEYTLVISNFNSQLSKVDLLLKLEGET
*R1_ML        1 ............................................................
*R2_4Arr      1 ............................................................
*R2_ML        1 ............................................................
*R3_KE        1 ............................................................
*R3_KE.       1 ............................................................
*S1_G3S3      1 ............................................................
*S1_CM        1 ................S...........................................
*S2_4Arr      1 ............................................................
*S2_CM        1 ............................................................

*R1_L3-5     61 DNGLSVLNVTKMVDVRRNMNRMINFNMPEDLTAGNYKITIDGQRGFSFHKEAELVYLSKS
*R1_ML       61 ............................................................
*R2_4Arr     61 ............................................................
*R2_ML       61 ............................................................
*R3_KE       61 ............................................................
*R3_KE.      61 ............................................................
*S1_G3S3     61 ............................................................
*S1_CM       61 ..................I.......E..................................
*S2_4Arr     61 ............................................................
*S2_CM       61 ............................................................

*R1_L3-5    121 ISGLIQVDKPVFKPGDTVNFRVIVLDTELKPPARVKSVYVTIRDPQRNVIRKWSTAKLYA
*R1_ML      121 ............................................................
*R2_4Arr    121 ............................................................
*R2_ML      121 ............................................................
*R3_KE      121 ............................................................
*R3_KE.     121 ..................................H.........................
*S1_G3S3    121 ..................................H.........................
*S1_CM      121 ............................................................
*S2_4Arr    121 ............................................................
*S2_CM      121 ............................................................

*R1_L3-5    181 GVFESDLQIAPTPMLGVWNISVEVEGEELVSKTFEVKEYVLSTFDVQVMPSVIPLEEHQA
*R1_ML      181 ............................................................
*R2_4Arr    181 ............................................................
*R2_ML      181 ............................................................
*R3_KE      181 ............................................................
*R3_KE.     181 ............................................................
*S1_G3S3    181 ............................................................
*S1_CM      181 ............................................................
*S2_4Arr    181 ............................................................
*S2_CM      181 ............................................................
                                    E266, N289, MG3
*R1_L3-5    241 VNLTIEANYHFGKPVQGVAKVELYLDDDKLKLKKELTVYGKGQVELRFDNFAMDADQQDV
*R1_ML      241 ............................................................
*R2_4Arr    241 .....D.....................NQ..........................
*R2_ML      241 .....D.....................NQ..........................
*R3_KE      241 .T.....................E....NQ..............N............
*R3_KE.     241 .....................E....NQ..............N............
*S1_G3S3    241 ............................NQ..........................
*S1_CM      241 ............................NQ..........................
*S2_4Arr    241 ............................NQ..........................
*S2_CM      241 ............................NQ..........................

*R1_L3-5    301 PVKVSFVEQYTNRTVVKQSQITVYRYAYRVELIKESPQFRPGLPFKCALQFTHHDGTPAK
*R1_ML      301 ......I.....................................................
*R2_4Arr    301 R.....I..H.................................................
*R2_ML      301 R.....I..H.................................................
*R3_KE      301 R.....I..H.................................................
*R3_KE.     301 R.....I..H.................................................
*S1_G3S3    301 R.....I.....................................................
*S1_CM      301 R.....I.....................................................
*S2_4Arr    301 R.....I.....................................................
*S2_CM      301 R.....I.....................................................
```


The full-length sequences of *TEP1*R3* (highlighted in grey color) compared to **R1* and **R2* alleles from Mali, **R2* and **S1* alleles from Cameroon, and **R1,*R2, *S1* alleles from insectary strains. Regions with amino acid modifications for *TEP1*R3* are highlighted in Turquoise color, and their positions give above the alignment.

```
*R1_L3-5  361 GISGKVEVSDVRFETTTTSDNDGLIKLELQPSEGTEQLSIHFNAVDGFFFYEDVNKVETV
*R1_ML    361 ..........................................................N....
*R2_4Arr  361 ..T........G.........................G.N...................
*R2_ML    361 ..T........G.........................G.N...................
*R3_KE    361 ...........G.........................G.S...................
*R3_KE.   361 ...........G.........................G.S...................
*S1_G3S3  361 ..T........G....K.................S...G.N...................
*S1_CM    361 ..T........G.........................G.N...................
*S2_4Arr  361 ..T........G.........................G.N...................
*S2_CM    361 ..T........G.........................G.N...................

*R1_L3-5  421 TDAYIKLELKSPIKRNKLMRFMVTCTERMTFFVYYVMSKGNIIDAGFMRPNKQPKYLLQL
*R1_ML    421 .N......................................................T.....
*R2_4Arr  421 ..........................................................T.....
*R2_ML    421 ..........................................................T.....
*R3_KE    421 ..........................................................T.....
*R3_KE.   421 ..........................................................T.....
*S1_G3S3  421 ..........................................................T.....
*S1_CM    421 ..........................................................T.....
*S2_4Arr  421 ..........................................................T.....
*S2_CM    421 ..........................................................T.....

*R1_L3-5  481 NATEKMIPRAKILIATVAGRTVVYDFADLDFQELRNNFDLSIDEQEIKPGRQIELSMSGR
*R1_ML    481 ..........................................................F..
*R2_4Arr  481 ........K.....................................................
*R2_ML    481 ........K.....................................................
*R3_KE    481 ........K.....................................................
*R3_KE.   481 ........M.....................................................
*S1_G3S3  481 ........K.................Y.....................................
*S1_CM    481 ........K.................Y.....................................
*S2_4Arr  481 ........K.................Y.....................................
*S2_CM    481 ........K.................Y.....................................

*R1_L3-5  541 PGAYVGLAAYDKALLLFNKNHDLFWEDIGQVFDGFHAINENEFDIFHSLGLFARTLDDIL
*R1_ML    541 ..........................................................
*R2_4Arr  541 ..........................................................
*R2_ML    541 ..........................................................
*R3_KE    541 ..........................................................
*R3_KE.   541 ..........................................................
*S1_G3S3  541 ..........................................................
*S1_CM    541 ..........................................................
*S2_4Arr  541 ..........................................................
*S2_CM    541 ..........................................................

*R1_L3-5  601 FDSANEKTGRNALQSGKPIGKLVSYRTNFQESWLWKNVSIGRSGSRKLIEVVPDTTTSWY
*R1_ML    601 ..........................................................
*R2_4Arr  601 ..........................................................
*R2_ML    601 ..........................................................
*R3_KE    601 ..........................................................
*R3_KE.   601 ..........................................................
*S1_G3S3  601 ..........................................................
*S1_CM    601 ..........................................................
*S2_4Arr  601 ..........................................................
*S2_CM    601 ..........................................................

*R1_L3-5  661 LTGFSIDPVYGLGIIKKPIQFTTVQPFYIVENLPYSIKRGEAVVLQFTLFNNLGAEYIAD
*R1_ML    661 ..........................................................
*R2_4Arr  661 ..........................................................
*R2_ML    661 ..........................................................
*R3_KE    661 ..........................................................
*R3_KE.   661 ..........................................................
*S1_G3S3  661 ..........................................................
*S1_CM    661 ..........................................................
*S2_4Arr  661 ..........................................................
*S2_CM    661 ..........................................................
```


The full-length sequences of *TEP1*R3* (highlighted in grey color) compared to **R1* and **R2* alleles from Mali, **R2* and **S1* alleles from Cameroon, and **R1,*R2, *S1* alleles from insectary strains. Regions with amino acid modifications for *TEP1*R3* are highlighted in Turquoise color, and their positions give above the alignment.



The full-length sequences of *TEP1*R3* (highlighted in grey color) compared to **R1* and **R2* alleles from Mali, **R2* and **S1* alleles from Cameroon, and **R1,*R2, *S1* alleles from insectary strains. Regions with amino acid modifications for *TEP1*R3* are highlighted in Turquoise color, and their positions give above the alignment.

```
*R1_L3-5 1081 VMAEKYTDGIPVMNWLVNQRYVTGSFPSTQDTFVGLKALTKMAEKISPSRNDYTVQLKYK
*R1_ML   1081 ..............................................................
*R2_4Arr 1081 ..............................................................
*R2_ML   1081 ..............................................................
*R3_KE   1081 ......L.......................................................
*R3_KE.  1081 ......L.......................................................
*S1_G3S3 1081 ......L....I.............R.............L.....................
*S1_CM   1081 ......L.................R.............L.....................
*S2_4Arr 1081 ......L.................R.............L.....................
*S2_CM   1081 ......L.................R.............L.....................

*R1_L3-5 1141 KSAKYFKINSEQIDVENFVDIPEDTKKLEINVGGIGFGLLEVVYQFNLNLVNFENRFQLD
*R1_ML   1141 .C................G.........................D.............
*R2_4Arr 1141 ................G.....................................
*R2_ML   1141 ................G.....................................
*R3_KE   1141 ................L.........................D.............
*R3_KE.  1141 ................L.........................YD............
*S1_G3S3 1141 ..T...N.......FQ..LE....................I...D.......H..K..
*S1_CM   1141 ..I...N.......FQ..LE..Q.................I...D.......H..K..
*S2_4Arr 1141 ..T...N........Q..LE....................I...D.......H..K..
*S2_CM   1141 ..T...N........Q..LE....................I...D.......H..K..
                                  K1223 MG8
*R1_L3-5 1201 LEKQNTGSDYELRLKVCASYIPQLTDRRSNMALIEVTLPSGYVVDRNPISEQTKVNPIQK
*R1_ML   1201 ..............................................................
*R2_4Arr 1201 ..............................................................
*R2_ML   1201 ..............................................................
*R3_KE   1201 ..............R.......K.....................................
*R3_KE.  1201 ..............R.......K.............................T..L...
*S1_G3S3 1201 ..............R...N...E...SQ.........................T.....N
*S1_CM   1201 ..............R...N...E...SQ.........................T.....N
*S2_4Arr 1201 ..............R...N...E...SQ.........................T.....N
*S2_CM   1201 ..............R...N...E...SQ.........................T.....N

*R1_L3-5 1261 TEIRYGGTSVVLYYDNMGSERNCFTLTAYRRFKVALKRPAYVVVYDYYNTNLNAIKVYEV
*R1_ML   1261 ..............................................................
*R2_4Arr 1261 ..............................................................
*R2_ML   1261 ..............................................................
*R3_KE   1261 ..............................................................
*R3_KE.  1261 ..............N..........M.....................D...........
*S1_G3S3 1261 M.............Y...T......V..................................
*S1_CM   1261 M.............Y...T......V..................................
*S2_4Arr 1261 M.............Y...T......V..................................
*S2_CM   1261 M.............Y...T......V..................................

*R1_L3-5 1321 DKQNLCEICDEEDCPAEC
*R1_ML   1321 ..................
*R2_4Arr 1321 ..................
*R2_ML   1321 ..................
*R3_KE   1321 ..................
*R3_KE.  1321 ..................
*S1_G3S3 1321 ....V....E........
*S1_CM   1321 ....V....E........
*S2_4Arr 1321 ....V....E........
*S2_CM   1321 ....V....E........
```


The full-length sequences of *TEP1*R3* (highlighted in grey color) compared to **R1* and **R2* alleles from Mali, **R2* and **S1* alleles from Cameroon, and **R1,*R2, *S1* alleles from insectary strains. Regions with amino acid modifications for *TEP1*R3* are highlighted in Turquoise color, and their positions give above the alignment.

Appendix 7. R script used to assess the infections of *P. falciparum*

Assessing the infections of *P. falciparum*

A *P. falciparum* infection intensity on *H3T1* mosquitoes

```
# define custom median function
plot.median <- function(x) {
  m <- median(x)
  c(y = m, ymin = m, ymax = m)
}
# function for number of observations
give.n <- function(x){
  return(c(y = median(x)*1.05, label = length(x)))
  # experiment with the multiplier to find the perfect position
}#usage: +stat_summary(fun.data = give.n, geom = "text", fun.y = median, colour="red")
# function for median labels
median.n <- function(x){
  return(c(y = median(x)*0.97, label = round(median(x),2)))
  # experiment with the multiplier to find the perfect position
}#usage: +stat_summary(fun.data = median.n, geom = "text", fun.y = median, colour = "red")
# function for mean labels
mean.n <- function(x){
  return(c(y = median(x)*0.97, label = round(mean(x),2)))
  # experiment with the multiplier to find the perfect position
}#usage: +stat_summary(fun.data = mean.n, geom = "text", fun.y = median, colour = "red")
# function for mean labels
#Easiest way of summarizing data using summarySE function together with a plyr package
summarySE <- function(data=NULL, measurevar, groupvars=NULL, na.rm=FALSE,
                      conf.interval=.95, .drop=TRUE) {
  library(plyr)

  # New version of length which can handle NA's: if na.rm==T, don't count them
  length2 <- function (x, na.rm=FALSE) {
    if (na.rm) sum(!is.na(x))
    else       length(x)
  }
  # This does the summary. For each group's data frame, return a vector with
  # N, mean, and sd
  datac <- ddply(data, groupvars, .drop=.drop,
    .fun = function(xx, col) {
      c(N    = length2(xx[[col]], na.rm=na.rm),
        mean = mean   (xx[[col]], na.rm=na.rm),
        sd   = sd     (xx[[col]], na.rm=na.rm)
      )
    },
    measurevar
```

Assessing the infections of *P. falciparum*

```
            )
            # Rename the "mean" column
    datac <- rename(datac, c("mean" = measurevar))
    datac$se <- datac$sd / sqrt(datac$N)  # Calculate standard error of the mean
    # Confidence interval multiplier for standard error
    # Calculate t-statistic for confidence interval:
    # e.g., if conf.interval is .95, use .975 (above/below), and use df=N-1
    ciMult <- qt(conf.interval/2 + .5, datac$N-1)
    datac$ci <- datac$se * ciMult
    return(datac)
}#Usage: data2 <- summarySE(data, measurevar="xxxx", groupvars=c("yyy",
"zzz"),na.rm=TRUE)
 #loading the data##read the csv file(MSDOS CSV). define the seperator & decimal sign
infdat <- read.csv ("MalvecBlokInfections.csv",sep=",",dec=".",na.strings=c("","NA"))
#for ploting
## define custom median function
plot.median <- function(x) {
  m <- median(x)
  c(y = m, ymin = m, ymax = m)
}
#Adding a new column "Infected" to show infection status
infdat$Infected[as.numeric(infdat$Oocyst) > 0] <- "Yes"
infdat$Infected[infdat$Oocyst == 0] <- "No"
#cleaning the data
infdat <- subset(infdat, !Genotype=="NEG")
infdat <- subset(infdat, !Genotype=="-")
infdat <- subset(infdat, !Oocyst=="-")
infdat <- subset(infdat, !Genotype=="NA")
infdat <- subset(infdat, !Oocyst=="NA")
infdat <- subset(infdat, !Genotype == "<NA>")
infdat <- subset(infdat, !Oocyst == "<NA>")
infdat <- subset(infdat, !Genotype=="")
infdat <- subset(infdat, !Genotype=="")
head(infdat)#confirm the data from the head
str(infdat)#check the structure of the data
summary(infdat$Infected)
summary(infdat$Genotype)
#getting the NF54 set for analyses, and zeroing into experiment 1
# Experiment 3 to 7 tests on influence on experiments o data sets
NF54dat <- subset(infdat, Parasite=="NF54")
NF54dat2to7<- subset(NF54dat, !Experiment== 1)
NF54dat3to7<- subset(NF54dat2to7, !Experiment== 2)
NF54dat$Oocyst<-as.numeric(as.character(NF54dat$Oocyst))
NF54dat3to7i<- subset(NF54dat, !Oocyst==0)#Exclude non-infected mosquitoes
#NF54dat3to7i<- subset(NF54dat)
#Infection intensity of NF54-experiments 3 to 7
NF54dat3to7i$Genotype <- factor(NF54dat3to7i$Genotype)
NF54dat3to7i$Oocyst <- as.numeric(as.charscter(NF54dat3to7i$Oocyst))
str(NF54dat3to7i)
```

Assessing the infections of *P. falciparum*

```
####Read about data transformation of zero values: https://stat.ethz.ch/pipermail/r-sig-ecology/2009-June/000676.html
NF54dat3to7$Oocyst<-as.numeric(as.character(NF54dat3to7$Oocyst))
quantile(NF54dat3to7$Oocyst, .0) #get zero quartile= 0
quantile(NF54dat3to7$Oocyst, .25) #get 1st quartile = 2
quantile(NF54dat3to7$Oocyst, .50) #get 2nd quartile = 9
quantile(NF54dat3to7$Oocyst, .75) #get 3rd quartile = 9
quantile(NF54dat3to7$Oocyst, 1.0) #get 4rd quartile = 25
#1. Eliminator (get rid of zeros). This way of eliminating zeros compromises on my phenotype of R1-allele bearing mosquitoes
#2. Oner (log1p)#Inbult in RI but ones needs to understand how it works
#3. Little-better (half of the smallest non-zero value). 2 way of adding half the lowest value e.g. in my case 1, is not convincing to our case as the data are skewed
#4. Quantiler (ratio of squared first and third quantile) log transformation by using squared 1st quartle divided by the third quartile DOES NOT WORK or apply in this case in this case.
#Conclusion. Either to include zeros and not to do log transformation for these analyses. However, its better to exclude oocyst ==0, in the assessment of infection loads, but include them when assessing the prevalence of infection to avoid bias in the data analyses, as has may have little effect on the phenotypes for (more) resistant genotypes e.g. *R1 alleles
#verify normality of the data
#using Q-Q plot
n=(NF54dat3to7i$Oocyst)
qqnorm(n)
abline(mean(n),sd(n))
#using shapiro test
shapiro.test(n)
#oocyst data set is not normally distributed therefore non-parametric statistical analyses is done on the infection loads.
k.w=kruskal.test(NF54dat3to7i$Oocyst~NF54dat3to7i$Genotype)
pairwise.wilcox.test(NF54dat3to7i$Oocyst,NF54dat3to7i$Genotype,p.adj="bonferroni")#there is clear effect of R1 bearing genotypes on infection loads compared to S1 and S2-bearing mosquitoes
#run parametric test on transformed data
NF54dat3to7i$Logload<-log(NF54dat3to7i$Oocyst)#log transform for parametric statistical analyses
av=aov(NF54dat3to7i$Logload~NF54dat3to7i$Genotype)
summary(av)
TukeyHSD(av)#there are NO significant differences between treatments i.e. influence of genotypes (alleles) on infection load)
#Plotting global infection intensity, Oocyst>=0
Nfload <- ggplot(NF54dat3to7i, aes(colour = factor(Genotype), x=Genotype, y = Oocyst))+geom_jitter(position = position_jitter(width = 0.9, height = 0.1))
NF54dat3to7i$Oocyst
Nfload <- Nfload + scale_colour_manual(values=c("#0072B2", "#999999", "#000000", "#009E73", "#E69F00", "#B2DF8A"),name="Genotype")+
  theme(legend.position="none")+
  theme(axis.text=element_text(size=10),axis.title=element_text(size=12)) + labs(x
```

Assessing the infections of *P. falciparum*

```
= "TEP1 genotype", y = "PfNF54 oocyst\nper midgut (Log10)")
    Nfload<- Nfload + theme(plot.title = element_text(size = rel(1))) + labs(title = "") +
    stat_summary(fun.data="plot.median", geom="errorbar", colour="red", width=0.5,
size=1)+scale_y_log10()#+    stat_summary(fun.data = give.n, geom = "text", fun.y =
median, colour ="red")
    Nfload#Fig. 3-5A
```

B Prevalence of *P. falciparum* infection intensity in *H3T1* mosquitoes

```
#experiment 3 (Note experiments 1 to 2 did not have R1/R1 genotypes so they were
excluded from these analyses)
NF54dat <- subset(infdat, Parasite=="NF54")
NF54dat3<- subset(NF54dat, Experiment== 3)
NF54dat3$Oocyst<-as.numeric(as.character(NF54dat3$Oocyst))
plot(NF54dat3$Oocyst~as.factor(NF54dat3$Genotype))
NF54dat3$Experiment<-drop.levels(NF54dat3$Experiment)
NF54dat3$Oocyst
mytable3 <- xtabs(~Infected+Genotype, data=NF54dat3, drop.unused.levels = T)
mytable3
NF54dat3$Genotype <- ordered(NF54dat3$Genotype, levels=c("R1/R1", "R1/S1",
"R1/S2","S1/S1","S1/S2","S2/S2"))
#Infection resistance in percentage
NF54dat3$Genotype <- factor(NF54dat3$Genotype)
str(NF54dat3)
n3<- table(NF54dat3$Genotype,NF54dat3$Infected)
rownames(n3)
colnames(n3)
addmargins(n3)
Row_total <- n3[,1]+n3[,2]
row.names <- factor(c("R1/R1", "R1/S1", "R1/S2", "S1/S1","S1/S2","S2/S2"))
Resistance <- (n3[,1]/Row_total)*100
Prevalence <- (n3[,2]/Row_total)*100
Result_table3 <- as.data.frame(cbind(n3,Row_total,Resistance,Prevalence))
str(Result_table3)
Result_table3$Genotype <- factor(c("R1/R1", "R1/S1", "R1/S2",
"S1/S1","S1/S2","S2/S2"))
Result_table3$Experiment <- factor(3)
t3f$Resistance<-as.numeric(t3f$Resistance)
#NF54 experiment 4
NF54dat4<- subset(NF54dat, Experiment== 4)
NF54dat4$Oocyst<-as.numeric(as.character(NF54dat4$Oocyst))
#Infection resistance in percentage
NF54dat4$Genotype <- factor(NF54dat4$Genotype)
n4<- table(NF54dat4$Genotype,NF54dat4$Infected)
Row_total <- n4[,1]+n4[,2]
row.names <- factor(c("R1/R1", "R1/S1", "R1/S2", "S1/S1","S1/S2","S2/S2"))
Resistance <- (n4[,1]/Row_total)*100
Prevalence <- (n4[,2]/Row_total)*100
Result_table4 <- as.data.frame(cbind(n4,Row_total,Resistance,Prevalence))
Result_table4$Genotype <- factor(c("R1/R1", "R1/S1", "R1/S2",
"S1/S1","S1/S2","S2/S2"))
```

Assessing the infections of *P. falciparum*

```
Result_table4$Experiment <- factor(4)
t4f<-Result_table4
# NF54 Experiment 5
NF54dat5<- subset(NF54dat, Experiment== 5)
NF54dat5$Oocyst<-as.numeric(as.character(NF54dat5$Oocyst))
#Infection resistance in percentage
NF54dat5$Genotype <- factor(NF54dat5$Genotype)
n5<- table(NF54dat5$Genotype,NF54dat5$Infected)
Row_total <- n5[,1]+n5[,2]
Resistance <- (n5[,1]/Row_total)*100
Prevalence <- (n5[,2]/Row_total)*100
Result_table5 <- as.data.frame(cbind(n5,Row_total,Resistance,Prevalence))
Result_table5$Genotype <- factor(c("R1/S1", "R1/S2", "S1/S1","S1/S2","S2/S2"))
Result_table5$Experiment <- factor(5)
t5f<-Result_table5
# NF54 Experiments 6
NF54dat6<- subset(NF54dat, Experiment== 6)
NF54dat6$Oocyst<-as.numeric(as.character(NF54dat6$Oocyst))
mytable6 <- xtabs(~Infected+Genotype, data=NF54dat6, drop.unused.levels = T)
NF54dat6$Genotype <- ordered(NF54dat6$Genotype, levels=c("R1/S1",
"R1/S2","S1/S1","S1/S2","S2/S2"))
#Infection resistance in percentage
NF54dat6$Genotype <- factor(NF54dat6$Genotype)
n6<- table(NF54dat6$Genotype,NF54dat6$Infected)
Row_total <- n6[,1]+n6[,2]
Resistance <- (n6[,1]/Row_total)*100
Prevalence <- (n6[,2]/Row_total)*100
Result_table6 <- as.data.frame(cbind(n6,Row_total,Resistance,Prevalence))
Result_table6$Genotype <- factor(c("R1/S1", "R1/S2", "S1/S1","S1/S2","S2/S2"))
Result_table6$Experiment <- factor(6)
t6f<-Result_table6
# NF54 Experiment 7
NF54dat7<- subset(NF54dat, Experiment== 7)
NF54dat7$Oocyst<-as.numeric(as.character(NF54dat7$Oocyst))
NF54dat7$Genotype <- factor(NF54dat7$Genotype)
n7<- table(NF54dat7$Genotype,NF54dat7$Infected)
Row_total <- n7[,1]+n7[,2]
row.names <- factor(c("R1/R1", "R1/S1", "S1/S1","S1/S2","S2/S2"))
Resistance <- (n7[,1]/Row_total)*100
Prevalence <- (n7[,2]/Row_total)*100
Result_table7 <- as.data.frame(cbind(n7,Row_total,Resistance,Prevalence))
Result_table7$Genotype <- factor(c("R1/R1", "R1/S1",  "S1/S1","S1/S2","S2/S2"))
Result_table7$Experiment <- factor(7)
t7f<-Result_table7
#Bind the summary tables for t3f to t7 for doing the SE summary statistics
library(dplyr)
t3to7<-bind_rows(list(t3f, t4f,t5f,t6f,t7f))
#t3to7<-bind_rows(list(t3,t5f,t6,t7))
t3to7
```

Assessing the infections of *P. falciparum*

```
detach("package:dplyr")#functions in dplyr masked those of plyr useful for running the summarrySE function for standard error, so I keep detaching it after use
#Statistics on prevalence and resistance
t3to7s<-t3to7
#t3to7s$Prevalence<-round(t3to7s$Prevalence,digits=0)
#t3to7s$Resistance<-round(t3to7s$Resistance,digits=0)
t3to7s$Prevalence<-as.numeric(as.character(t3to7s$Prevalence))
t3to7s$Resistance<-as.numeric(as.character(t3to7s$Resistance))
t3to7s$Genotype<-factor(t3to7s$Genotype)
names(t3to7s)
#Verify normality of the data using Q-Q plot and shapiro test
n=(t3to7s$Resistance)
qqnorm(n)
abline(mean(n),sd(n))
shapiro.test(t3to7s$Resistance)
#My data is normally distributed so I will use parametric analyses to compare their medians
#will not Run non-parametric tests
k.w=kruskal.test(t3to7s$Resistance~t3to7s$Genotype)
k.w
pairwise.wilcox.test(t3to7s$Resistance,t3to7s$Genotype,p.adj="bonferroni")
#Will run parametric test
av=aov(t3to7s$Resistance~t3to7s$Genotype)
summary(av)
TukeyHSD(av)
n=(t3to7s$Prevalence)
qqnorm(n)
abline(mean(n),sd(n))
shapiro.test(t3to7s$Prevalence)
#My data is normally distributed so I will use parametric analyses to compare their means#will not Run non-parametric tests e.g.
#k.w=kruskal.test(t3to7s$Prevalence~t3to7s$Genotype)
#pairwise.wilcox.test(t3to7s$Prevalence,t3to7s$Genotype,p.adj="bonferroni")
#But will run parametric test instead
av=aov(t3to7s$Prevalence~t3to7s$Genotype)
summary(av)
TukeyHSD(av)#No significant differences
#Convert groups into factor
t3to7$Genotype<- factor(t3to7$Genotype)
t3to7$Experiment<- factor(t3to7$Experiment)
t3to7$Resistance<- as.numeric(as.character(t3to7$Resistance))
#Then sumarizing the data using the above summarySE function
nfbd3to7 <- summarySE(t3to7, measurevar="Resistance", groupvars="Genotype",na.rm=TRUE)
nfbd3to7p <- summarySE(t3to7, measurevar="Prevalence", groupvars=c("Genotype"),na.rm=TRUE)
# Rename column Resistance and Prevalence to just Resistance.mean and prevalence.mean respectively
names(nfbd3to7)[names(nfbd3to7)=="Resistance"] <- "Resistance.mean"
```

Assessing the infections of *P. falciparum*

```
names(nfbd3to7p)[names(nfbd3to7p)=="Prevalence"] <- "Prevalence.mean"
#used data frame t3to7 to do statistics on resistance or prevalence
# Define the top and bottom of the errorbars
limits3to7 <- aes(ymax = Resistance.mean + se, ymin=Resistance.mean - se)#for resistance
limits3to7p <- aes(ymax = Prevalence.mean + se, ymin=Prevalence.mean - se)#for prevalence
#Begin your ggplot
#Here we are plotting Experiment vs mean of resistance and filling by another factor grouped variable Genotype
nf3to7<-ggplot(nfbd3to7,aes(Genotype,Resistance.mean,fill=Genotype))
#Creating bar to show the factor variable position_dodge
#ensures side by side creation of factor bars
nf3to7<-nf3to7+geom_bar(stat = "identity",position = position_dodge())
#creation of error bar
nf3to7<-nf3to7+geom_errorbar(limits3to7,width=0.25,position = position_dodge(width = 0.9))
#Making the final plot
nf3to7<-nf3to7+ggtitle ("PfNF54 overall resistance") +
  xlab ("TEP1 genotype") + ylab("Mean  % resistant mosquitoes") +
  theme(legend.title = element_text(colour="black", size=10)) +ylim(c(0,100))+
  theme(legend.text = element_text(colour="black", size=10, face="italic"))+scale_fill_grey(start = 0.3, end = 0.8,na.value = "red")+theme(legend.position="none")
#+scale_fill_manual(values=c("#0072B2", "#999999", "#000000","#009E73", "#E69F00", "#B2DF8A"),name="Genotype",labels=c("R1/R1", "R1/S1", "R1/S2", "S1/S1", "S1/S2","S2/S2"))
nf3to7
#Prevalence: Here we are plotting Experiment vs mean of prevalence and filling by another factor variable Genotype
nf3to7p<-ggplot(nfbd3to7p,aes(Genotype,Prevalence.mean, fill=Genotype))
#Creating bar to show the factor variable position_dodge
#ensures side by side creation of factor bars
nf3to7p<-nf3to7p+geom_bar(stat = "identity",position = position_dodge())
#creation of error bar
nf3to7p<-nf3to7p+geom_errorbar(limits3to7p,width=0.25,position = position_dodge(width = 0.9))
#Making the final plot
nf3to7p<-nf3to7p+ggtitle ("") +
  xlab ("TEP1 genotype") + ylab("Mean  % PfNF54 prevalence") +
  theme(legend.title = element_text(colour="black", size=10)) +ylim(c(0,100))+
  theme(legend.text = element_text(colour="black", size=10, face="italic"))+scale_fill_grey(start = 0.3, end = 0.8,na.value = "red")+theme(legend.position="none")
#+scale_fill_manual(values=c("#0072B2", "#999999", "#000000","#009E73", "#E69F00", "#B2DF8A"),name="Genotype",labels=c("R1/R1", "R1/S1", "R1/S2", "S1/S1", "S1/S2","S2/S2"))
nf3to7p#Fig. 3.5B
```

Appendix 8. Author's Curriculum Vitae (CV)

Removed for online publication

Submitted manuscript, reviewed and is under consideration

- M Gildenhard*, **EK Rono***, A Boissière, A Diarra, P Bascunan, D Camara, H Krüger, M Mariko, R Mariko, MK Rono, P Mireji, J Pompon, PB Seda, Y Thailayil, A Traorè, S. Nsango, P Awono-Ambene, RK Dabire, A Diabate, M Diallo, D Masiga, D Sangare, F Catteruccia, I Morlais, and EA Levashina. 'Mosquito microevolution drives *Plasmodium falciparum* dynamics'. *These authors equally contributed to this work. The manuscript was submitted to the Nature Journal on June 12, 2017.

Publication

- My 2012 Master thesis:
  Juliette Rose Ongus, **Evans Kiplangat Rono** and Fred Alexander Wafula Wamunyokoli. Silencing of the rift valley fever virus S-genome segment transcript using RNA interference in *Sf*21 insect cells. African Journal of Biotechnology (2017). Pgs. 1016-1031, Vol. 16 (18). DOI: 10.5897/AJB2017.15910.

List of published abstracts

1) **Evans K. Rono**, MALVECBLOK Consortium and Elena A. Levashina (May 2015). *TEP1* polymorphism in Africa. EMBL Conference BioMalPar XI: 'Biology and Pathology of the Malaria Parasite', Heldelberg, Germany. Pg 99.
2) **Evans K. Rono**, Masiga, D., Wamunyokoli, F., Gikunju, J. K., Sang, R. and Masiga, D., Ongus, J. (Nov 16, 2011). JKUAT 6th Scientific and industrialization conference, Kenya, Nairobi. Title of poster presentation

'Susceptibility of Rift Valley Fever Virus S Genome RNA Transcripts to RNA Interference'. Pg 43.

3) **Evans K. Rono,** Wamunyokoli, F., Gikunju, J. K., Sang, R., Masiga, D. and Ongus, J. (Sep 8-9th 2011). The Baculovirus Expression Vector System as A Model for Demonstrating the Susceptibility of Rift Valley Fever Virus (RVFV) S-Genome Segment to RNA Silencing. 1st Medical and Veterinary Virus Research, Kenya, Nairobi. Pg 7.

List of talks and/or poster presentations, conferences, meetings and workshops

1) **Evans K. Rono,** Markus Gildenhard, MALVECBLOK Consortium and Elena A. Levashina (May, 2016). *TEP1* allelic variation in shaping vector population and in malaria transmission. Max Planck Institute for Infection Biology (MPIIB) 8th Scientific Retreat, Hotel Döllnsee-Schorfheide, Germany.
2) **Talk.** September. 2015: Strasbourg- Berlin LIA meeting, Berlin, Germany. Title "*TEP1* polymorphism and development of *Plasmodium falciparum*".
3) **Evans K. Rono**, Masiga, D., Wamunyokoli, F., Gikunju, J. K., Sang, R. and Masiga, D., Ongus, J. (Nov 16, 2011). *'Susceptibility of Rift Valley Fever Virus S Genome RNA Transcripts to RNA Interference'.* At *icipe,* Kenya, Nairobi.
4) **Evans K. Rono,** Wamunyokoli, F., Gikunju, J. K., Sang, R., Masiga, D. and Ongus, J. (Sep 8-9th 2011). The Baculovirus Expression Vector System as A Model for Demonstrating the Susceptibility of Rift Valley Fever Virus (RVFV) S-Genome Segment to RNA Silencing'. 1st Medical and Veterinary Virus Research, Kenya, Nairobi.
5) June 2016: ZIBI Summer Symposium, Berlin, Germany.
6) May, 2016: Max Planck Institute for Infection Biology, 8th Scientific Retreat, Hotel Döllnsee-Schorfheide, Germany.
7) September, 2015: Strasbourg- Berlin LIA meeting, Berlin, Germany. Title "*TEP1* polymorphism and development of *Plasmodium falciparum*".
8) June, 2015: ZIBI Summer Symposium: 'New concepts in pathogen sensing and host defense', Berlin, Germany.
9) May, 2015: EMBL Conference BioMalPar XI: 'Biology and Pathology of the Malaria Parasite', Heldelberg, Germany. Title *'TEP1 polymorphism In Africa'.*
10) May, 2015: MPIIB 7th Scientific Retreat, Berlin, Germany. Moderating a group discussion on *'Towards field and clinical research'.*
11) November, 2014: The DAAD Scholarship holders Meeting, Humboldt University, Berlin, Germany.
12) October, 2014: IEMID meeting, Berlin, Germany. Title *'TEP1 polymorphism in Africa'.*
13) September, 2014: MPIIB/DRFZ Mouse Course. Theoretical and practical parts for obtaining "Fachkenntnis" for experiments with mice.
14) June, 2014: ZIBI Summer Symposium: 'Frontiers of Parasitology', Berlin, Germany.
15) May, 2014: Max Planck Institute for Infection Biology (MPIIB) 6th Scientific Retreat, Hotel Döllnsee-Schorfheide, Germany.
16) April, 2014: DAAD meeting on 'Public health', Würzburg, Germany.
17) February, 2014: Meeting of the CNRS-MPG International Associated Laboratory on 'REL2 and Resistance to Malaria', Strasbourg, France.
18) October, 2013: DAAD orientation meeting, Bonn, Germany.
19) June, 2012: Laboratory visit and meeting: Task: Discussion and interview on 'TEP1 polymorphism and PhD candidature' in Elena A. Levashina's laboratory at the Max-Planck Institute of Infection Biology. And to participated in The

ZIBI Summer Symposium 2012 "Global Challenges of Chronic Tropical Infections" in Berlin Germany. Organized by Humboldt-Universität zu Berlin; Center of Infection Biology and Immunity (ZIBI) Berlin, Germany.

20) April, 2012: Attended a 'Symposium 'Mosquito Biology and prospects for control of malaria in Africa' in Kenya. Organized by Elena A. Levashina and the MALVECBLOK project collaborators.
21) 2011: 1st Medical and Veterinary Virus Research, Kenya. Title *'Baculovirus Expression Vector System as A Model for Demonstrating the Susceptibility of Rift Valley Fever Virus (RVFV) S-Genome Segment to RNA Silencing'.*
22) 2011: Open Science Day during the 40th *icipe* Anniversary, Kenya. Title *'Susceptibility of Rift Valley Fever Virus S Genome RNA Transcripts to RNA Interference'.*
23) 2011: JKUAT 6th Scientific and industrialization conference, Kenya. Title *'Susceptibility of Rift Valley Fever Virus S Genome RNA Transcripts to RNA Interference'.*
24) 2011: Course on GIS: 'Introduction to Global Positioning System (GPS) and Global Positioning System (GPS) software for data sampling, visualization and analyses training at ICIPE, Kenya.

Reference List

1. WHO, "World Malaria Report," *World Health Organization* (2014).
2. WHO, "World Malaria Report," *World Health Organization* (2015).
3. M. E. Sinka *et al.*, The dominant Anopheles vectors of human malaria in Africa, Europe and the Middle East: occurrence data, distribution maps and bionomic precis. *Parasit Vectors* **3**, 117 (2010).
4. S. Blandin *et al.*, Complement-like protein TEP1 is a determinant of vectorial capacity in the malaria vector Anopheles gambiae. *Cell* **116**, 661-670 (2004).
5. S. A. Blandin *et al.*, Dissecting the Genetic Basis of Resistance to Malaria Parasites in Anopheles gambiae. *Science* **326**, 147-150 (2009).
6. D. J. Obbard *et al.*, The evolution of TEP1, an exceptionally polymorphic immunity gene in Anopheles gambiae. *BMC Evol. Biol.* **8**, (2008).
7. B. J. White *et al.*, Adaptive divergence between incipient species of Anopheles gambiae increases resistance to Plasmodium. *Proc Natl Acad Sci U S A* **108**, 244-249 (2011).
8. E. Mancini *et al.*, Adaptive Potential of Hybridization among Malaria Vectors: Introgression at the Immune Locus TEP1 between Anopheles coluzzii and A. gambiae in 'Far-West' Africa. *PLoS One* **10**, e0127804 (2015).
9. I. W. Sherman, in *Malaria Parasite Biology, Pathogenesis, and Protection,* I. W. Sherman, Ed. (American Society for Microbiology, United States of America, 1998), vol. 1, chap. 1, pp. 3-10.
10. K. M. Loban, E. S. Polozok, *Malaria*. A. Shelepin, Ed., (Mir Publishers, Union of Soviet Socialist Republics, 1989), pp. 260.
11. WHO, "World Malaria Report," *World Health Organization* (2013).
12. E. K. Rono, M. Gildenhard, C. MALVECBLOK, E. A. Levashina. (Unpublished work, 2016).
13. L. C. Gouagna *et al.*, Plasmodium falciparum malaria disease manifestations in humans and transmission to Anopheles gambiae: a field study in Western Kenya. *Parasitology* **128**, 235-243 (2004).
14. D. Gaur, D. C. Mayer, L. H. Miller, Parasite ligand-host receptor interactions during invasion of erythrocytes by Plasmodium merozoites. *Int. J. Parasitol.* **34**, 1413-1429 (2004).
15. J. W. Meigen, *Systematische beschreibung der bekannten europäischen zweiflügeligen insekten*. C. R. W. Wiedemann, H. Loew, Eds., (Halle: H.W. Schmidt, 1822).
16. P. L. Alonso *et al.*, A research agenda to underpin malaria eradication. *PLoS Med.* **8**, e1000406 (2011).
17. G. Volohonsky, S. Steinert, E. A. Levashina, Focusing on complement in the antiparasitic defense of mosquitoes. *Trends Parasitol.* **26**, 1-3 (2010).
18. E. A. Levashina *et al.*, Conserved role of a complement-like protein in phagocytosis revealed by dsRNA knockout in cultured cells of the mosquito, Anopheles gambiae. *Cell* **104**, 709-718 (2001).
19. S. K. Sreenivasamurthy *et al.*, A compendium of molecules involved in vector-pathogen interactions pertaining to malaria. *Malar. J.* **12**, 216 (2013).
20. K. Siekmans *et al.*, Community case management of malaria: a pro-poor intervention in rural Kenya. *Int Health* **5**, 196-204 (2013).
21. R. C. Smith, J. Vega-Rodriguez, M. Jacobs-Lorena, The Plasmodium bottleneck: malaria parasite losses in the mosquito vector. *Mem. Inst. Oswaldo Cruz* **109**, 644-661 (2014).

22. J. F. Hillyer, C. Barreau, K. D. Vernick, Efficiency of salivary gland invasion by malaria sporozoites is controlled by rapid sporozoite destruction in the mosquito haemocoel. *Int. J. Parasitol.* **37**, 673-681 (2007).
23. B. Singh *et al.*, A large focus of naturally acquired Plasmodium knowlesi infections in human beings. *Lancet* **363**, 1017-1024 (2004).
24. D. E. Neafsey *et al.*, Mosquito genomics. Highly evolvable malaria vectors: the genomes of 16 Anopheles mosquitoes. *Science* **347**, 1258522 (2015).
25. G. C. Lanzaro, Y. Lee, Speciation in Anopheles gambiae - The Distribution of Genetic Polymorphism and Patterns of Reproductive Isolation Among Natural Populations. *Anopheles Mosquitoes - New Insights into Malaria Vectors*, 173-196 (2013).
26. G. Davidson, Anopheles Gambiae, a Complex of Species. *Bull. World Health Organ.* **31**, 625-634 (1964).
27. R. H. Hunt, M. Coetzee, M. Fettene, The Anopheles gambiae complex: a new species from Ethiopia. *Trans. R. Soc. Trop. Med. Hyg.* **92**, 231-235 (1998).
28. M. Coetzee *et al.*, Anopheles coluzzii and Anopheles amharicus, new members of the Anopheles gambiae complex. *Zootaxa* **3619**, 246-274 (2013).
29. G. B. White, Anopheles bwambae sp.n., a malaria vector in the Semliki Valley, Uganda, and its relationships with other sibling species of the An.gambiae complex (Diptera: Culicidae). *Syst. Entomol.* **10**, 501-522 (1985).
30. R. E. Harbach, H. Townson, L. G. Mukwaya, T. Adeniran, Use of rDNA-PCR to investigate the ecological distribution of Anopheles bwambae in relation to other members of the An.gambiae complex of mosquitoes in Bwamba County, Uganda. *Med. Vet. Entomol.* **11**, 329-334 (1997).
31. J. Brunhes, G. LeGoff, B. Geoffroy, in *ANNALES DE LA SOCIETE ENTOMOLOGIQUE DE FRANCE*. (SOC ENTOMOLOGIQUE FRANCE 45 RUE BUFFON, 75005 PARIS, FRANCE, 1997), vol. 33, pp. 173-183.
32. Coluzzi M, Sabatini A, Petrarca V, D. M. A. Di, Chromosomal differentiation and adaptation to human environments in the Anopheles gambiae complex. *Trans. R. Soc. Trop. Med. Hyg* **73**, 483–497 (1979).
33. W. C. Black, G. C. Lanzaro, Distribution of genetic variation among chromosomal forms of Anopheles gambiae s.s.: introgressive hybridization, adaptive inversions, or recent reproductive isolation? *Insect Mol. Biol.* **10**, 3-7 (2001).
34. A. della Torre *et al.*, Speciation within Anopheles gambiae - the glass is half full. *Science* **298**, 115-117 (2002).
35. B. J. White *et al.*, Adaptive divergence between incipient species of Anopheles gambiae increases resistance to Plasmodium. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 244-249 (2011).
36. N. J. Besansky *et al.*, Semipermeable species boundaries between Anopheles gambiae and Anopheles arabiensis: Evidence from multilocus DNA sequence variation. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 10818-10823 (2003).
37. N. J. Besansky *et al.*, Molecular phylogeny of the Anopheles gambiae complex suggests genetic introgression between principal malaria vectors. *Proc. Natl. Acad. Sci. U. S. A.* **91**, 6885-6888 (1994).
38. M. C. Fontaine *et al.*, Mosquito genomics. Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science* **347**, 1258524 (2015).

39. C. D. Marsden *et al.*, Asymmetric introgression between the M and S forms of the malaria vector, Anopheles gambiae, maintains divergence despite extensive hybridization. *Mol. Ecol.* **20**, 4983-4994 (2011).
40. N. J. Thelwell, R. A. Huisman, R. E. Harbach, R. K. Butlin, Evidence for mitochondrial introgression between Anopheles bwambae and Anopheles gambiae. *Insect Mol. Biol.* **9**, 203-210 (2000).
41. A. Diabate *et al.*, Evidence for divergent selection between the molecular forms of Anopheles gambiae: role of predation. *BMC Evol. Biol.* **8**, (2008).
42. B. J. White *et al.*, Dose and developmental responses of Anopheles merus larvae to salinity. *J. Exp. Biol.* **216**, 3433-3441 (2013).
43. H. A. Smith *et al.*, Genome-wide QTL mapping of saltwater tolerance in sibling species of Anopheles (malaria vector) mosquitoes. *Heredity (Edinb.)*, (2015).
44. J. G. Halcrow, A new sub-species of Anopheles gambiae giles from Mauritius. *East Afr. Med. J.* **34**, 133-135 (1957).
45. C. Starr, in *Biology : concepts and applications,* C. A. Evers, L. Starr, Eds. (Stamford : Cengage Learning, Stamford, 2015), chap. 43, pp. 763-782.
46. Y. T. Toure *et al.*, The distribution and inversion polymorphism of chromosomally recognized taxa of the Anopheles gambiae complex in Mali, West Africa. *Parassitologia* **40**, 477-511 (1998).
47. A. Diabate *et al.*, Larval development of the molecular forms of Anopheles gambiae (Diptera: Culicidae) in different habitats: a transplantation experiment. *J. Med. Entomol.* **42**, 548-553 (2005).
48. A. Dao *et al.*, Signatures of aestivation and migration in Sahelian malaria mosquito populations. *Nature* **516**, 387-390 (2014).
49. M. Coetzee, Distribution of the African malaria vectors of the Anopheles gambiae complex. *Am. J. Trop. Med. Hyg.* **70**, 103-104 (2004).
50. G. B. White, The Anopheles gambiae complex and malaria transmission around Kisumu, Kenya. *Trans. R. Soc. Trop. Med. Hyg.* **66**, 572-581 (1972).
51. K. Murphy, in *Immunobiology,* P. Travers, M. Walport, C. A. Janeway, Eds. (New York, NY [u.a.] : Garland : New York, NY [u.a.] : Garland, New York, NY [u.a.], 2008), chap. 2, pp. 39-108.
52. I. Roitt, J. Brostoff, D. Male, in *Immunology,* L. Crowe, Ed. (Dianne Zack, Mosby International Limited, United Kingdom, 1998), chap. 4, pp. 43-60.
53. C. Speth, W. M. Prodinger, R. Würzner, H. Stoiber, M. P. Dierich, in *Fundamental Immunology,* W. E. Paul, Ed. (Lippincott Williams & Wilkins, United States of America, 2008), chap. 33, pp. 1047-1078.
54. M. Nonaka, Origin and evolution of the complement system. *Curr. Top. Microbiol. Immunol.* **248**, 37-50 (2000).
55. L. Sompayrac, in *How the immune system works*. (Chichester : Wiley-Blackwell, Chichester, 2016), chap. 2, pp. 13-26.
56. J. Vizioli *et al.*, Cloning and analysis of a cecropin gene from the malaria vector mosquito, Anopheles gambiae. *Insect Mol. Biol.* **9**, 75-84 (2000).
57. A. M. Richman, G. Dimopoulos, D. Seeley, F. C. Kafatos, Plasmodium activates the innate immune response of Anopheles gambiae mosquitoes. *EMBO J.* **16**, 6114-6119 (1997).
58. A. M. Richman *et al.*, Inducible immune factors of the vector mosquito Anopheles gambiae: biochemical purification of a defensin antibacterial peptide and molecular cloning of preprodefensin cDNA. *Insect Mol. Biol.* **5**, 203-210 (1996).

59. J. Vizioli *et al.*, Gambicin: a novel immune responsive antimicrobial peptide from the malaria vector Anopheles gambiae. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 12630-12635 (2001).
60. C. Luo, L. Zheng, Independent evolution of Toll and related genes in insects and mammals. *Immunogenetics* **51**, 92-98 (2000).
61. A. Goto, S. Blandin, J. Royet, J. M. Reichhart, E. A. Levashina, Silencing of Toll pathway components by direct injection of double-stranded RNA into Drosophila adult flies. *Nucleic Acids Res.* **31**, 6619-6623 (2003).
62. S. Meister *et al.*, Immune signaling pathways regulating bacterial and malaria parasite infection of the mosquito Anopheles gambiae. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 11420-11425 (2005).
63. L. S. Garver, G. de Almeida Oliveira, C. Barillas-Mury, The JNK pathway is a key mediator of Anopheles gambiae antiplasmodial immunity. *PLoS Pathog.* **9**, e1003622 (2013).
64. L. Gupta *et al.*, The STAT pathway mediates late-phase immunity against Plasmodium in the mosquito Anopheles gambiae. *Cell Host Microbe* **5**, 498-507 (2009).
65. G. K. Christophides *et al.*, Immunity-related genes and gene families in Anopheles gambiae. *Science* **298**, 159-165 (2002).
66. A. M. Mendes *et al.*, Infection intensity-dependent responses of Anopheles gambiae to the African malaria parasite Plasmodium falciparum. *Infect. Immun.* **79**, 4708-4715 (2011).
67. C. Barillas-Mury *et al.*, Immune factor Gambif1, a new rel family member from the human malaria vector, Anopheles gambiae. *EMBO J.* **15**, 4691-4701 (1996).
68. L. S. Garver, Y. Dong, G. Dimopoulos, Caspar controls resistance to Plasmodium falciparum in diverse anopheline species. *PLoS Pathog.* **5**, e1000335 (2009).
69. G. Dimopoulos, H. M. Muller, E. A. Levashina, F. C. Kafatos, Innate immune defense against malaria infection in the mosquito. *Curr. Opin. Immunol.* **13**, 79-88 (2001).
70. L. C. Gouagna *et al.*, The early sporogonic cycle of Plasmodium falciparum in laboratory-infected Anopheles gambiae: an estimation of parasite efficacy. *Trop. Med. Int. Health* **3**, 21-28 (1998).
71. S. A. Blandin, E. Marois, E. A. Levashina, Antimalarial responses in Anopheles gambiae: from a complement-like protein to a complement-like pathway. *Cell Host Microbe* **3**, 364-374 (2008).
72. M. M. Whitten, S. H. Shiao, E. A. Levashina, Mosquito midguts and malaria: cell biology, compartmentalization and immunology. *Parasite Immunol.* **28**, 121-130 (2006).
73. G. Dimopoulos, A. Richman, H. M. Muller, F. C. Kafatos, Molecular immune responses of the mosquito Anopheles gambiae to bacteria and malaria parasites. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 11508-11513 (1997).
74. L. C. Smith, L. Chang, R. J. Britten, E. H. Davidson, Sea urchin genes expressed in activated coelomocytes are identified by expressed sequence tags. Complement homologues and other putative immune response genes suggest immune system homology within the deuterostomes. *J. Immunol.* **156**, 593-602 (1996).
75. W. Z. Al-Sharif, J. O. Sunyer, J. D. Lambris, L. C. Smith, Sea urchin coelomocytes specifically express a homologue of the complement component C3. *J. Immunol.* **160**, 2983-2997 (1998).

76. B. V. Le, M. Williams, S. Logarajah, R. H. Baxter, Molecular basis for genetic resistance of Anopheles gambiae to Plasmodium: structural analysis of TEP1 susceptible and resistant alleles. *PLoS Pathog.* **8**, e1002958 (2012).
77. R. H. Baxter *et al.*, Structural basis for conserved complement factor-like function in the antimalarial protein TEP1. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 11615-11620 (2007).
78. G. Volohonsky *et al.*, Transgenic Expression of the Anti-parasitic Factor TEP1 in the Malaria Mosquito Anopheles gambiae. *PLoS Pathog.* **13**, e1006113 (2017).
79. E. A. Levashina, Immune responses in Anopheles gambiae. *Insect Biochem. Mol. Biol.* **34**, 673-678 (2004).
80. J. Pompon, E. A. Levashina, A New Role of the Mosquito Complement-like Cascade in Male Fertility in Anopheles gambiae. *PLoS Biol.* **13**, e1002255 (2015).
81. A. Oliveira Gde, J. Lieberman, C. Barillas-Mury, Epithelial nitration by a peroxidase/NOX5 system mediates mosquito antiplasmodial immunity. *Science* **335**, 856-859 (2012).
82. M. M. Riehle *et al.*, Natural malaria infection in Anopheles gambiae is regulated by a single genomic control region. *Science* **312**, 577-579 (2006).
83. M. Fraiture *et al.*, Two mosquito LRR proteins function as complement control factors in the TEP1-mediated killing of Plasmodium. *Cell Host Microbe* **5**, 273-284 (2009).
84. M. M. Riehle *et al.*, Anopheles gambiae APL1 is a family of variable LRR proteins required for Rel1-mediated protection from the malaria parasite, Plasmodium berghei. *PLoS One* **3**, e3672 (2008).
85. D. L. Hartl, A. G. Clark, in *Principles of population genetics,* D. L. Hartl, A. G. Clark, Eds. (Sunderland, Mass. : Sinauer : Sunderland, Mass. : Sinauer, Sunderland, Mass., 2007), chap. 1, pp. 3-43.
86. P. W. Hedrick, in *Genetics of Populations*. (Jones and Bartlett Publishers, USA, 2011), chap. 3, pp. 111-186.
87. M. Nei, F-statistics and analysis of gene diversity in subdivided populations. *Ann. Hum. Genet.* **41**, 225-233 (1977).
88. D. L. Harti, A. G. Clark, *Principles of Population Genetics*. (Sinauer Associates, USA, ed. 4, 2007), pp. 652.
89. W. H. Lowe, R. P. Kovach, F. W. Allendorf, Population Genetics and Demography Unite Ecology and Evolution. *Trends Ecol. Evol.* **32**, 141-152 (2017).
90. D. L. Hartl, A. G. Clark, in *Principles of population genetics,* D. L. Hartl, A. G. Clark, Eds. (Sunderland, Mass. : Sinauer : Sunderland, Mass. : Sinauer, Sunderland, Mass., 2007), chap. 2, pp. 45-93.
91. F. Singer, in *Ecology in action*. (Cambridge : Cambridge University Press, Cambridge, 2016), chap. 3, pp. 58-85.
92. M. W. Strickberger, in *Evolution*. (Sudbury, Mass. [u.a.] : Jones and Bartlett : Sudbury, Mass. [u.a.] : Jones and Bartlett, Sudbury, Mass. [u.a.], 2000), chap. 21, pp. 515-532.
93. M. W. Strickberger, in *Evolution*. (Sudbury, Mass. [u.a.] : Jones and Bartlett : Sudbury, Mass. [u.a.] : Jones and Bartlett, Sudbury, Mass. [u.a.], 2000), chap. 22, pp. 533-552.

94. D. L. Hartl, A. G. Clark, in *Principles of population genetics,* D. L. Hartl, A. G. Clark, Eds. (Sunderland, Mass. : Sinauer : Sunderland, Mass. : Sinauer, Sunderland, Mass., 2007), chap. 6, pp. 257-315.
95. D. L. Hartl, A. G. Clark, in *Principles of population genetics,* D. L. Hartl, A. G. Clark, Eds. (Sunderland, Mass. : Sinauer : Sunderland, Mass. : Sinauer, Sunderland, Mass., 2007), chap. 5, pp. 199-255.
96. D. J. Obbard, J. J. Welch, K. W. Kim, F. M. Jiggins, Quantifying adaptive evolution in the Drosophila immune system. *PLoS Genet.* **5**, e1000698 (2009).
97. S. J. McTaggart, D. J. Obbard, C. Conlon, T. J. Little, Immune genes undergo more adaptive evolution than non-immune system genes in Daphnia pulex. *BMC Evol. Biol.* **12**, 63 (2012).
98. D. Schluter, G. L. Conte, Genetics and ecological speciation. *Proc. Natl. Acad. Sci. U. S. A.* **106 Suppl 1**, 9955-9962 (2009).
99. A. E. Yawson, D. Weetman, M. D. Wilson, M. J. Donnelly, Ecological zones rather than molecular forms predict genetic differentiation in the malaria vector Anopheles gambiae s.s. in Ghana. *Genetics* **175**, 751-761 (2007).
100. A. Parmakelis *et al.*, The molecular evolution of four anti-malarial immune genes in the Anopheles gambiae species complex. *BMC Evol. Biol.* **8**, (2008).
101. D. J. Obbard, Y. M. Linton, F. M. Jiggins, G. Yan, T. J. Little, Population genetics of Plasmodium resistance genes in Anopheles gambiae: no evidence for strong selection. *Mol. Ecol.* **16**, 3497-3510 (2007).
102. F. Simard *et al.*, Ecological niche partitioning between Anopheles gambiae molecular forms in Cameroon: the ecological side of speciation. *BMC Ecol.* **9**, 17 (2009).
103. N. C. Manoukis *et al.*, A test of the chromosomal theory of ecotypic speciation in Anopheles gambiae. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 2940-2945 (2008).
104. M. Eldering *et al.*, Variation in susceptibility of African Plasmodium falciparum malaria parasites to TEP1 mediated killing in Anopheles gambiae mosquitoes. *Sci. Rep.* **6**, 20440 (2016).
105. R. S. McCann *et al.*, Reemergence of Anopheles funestus as a vector of Plasmodium falciparum in western Kenya after long-term implementation of insecticide-treated bed nets. *Am. J. Trop. Med. Hyg.* **90**, 597-604 (2014).
106. R. F. Beach *et al.*, Effectiveness of permethrin-impregnated bed nets and curtains for malaria control in a holoendemic area of western Kenya. *Am. J. Trop. Med. Hyg.* **49**, 290-300 (1993).
107. G. Zhou *et al.*, Modest additive effects of integrated vector control measures on malaria prevalence and transmission in western Kenya. *Malar. J.* **12**, 256 (2013).
108. F. C. Tanser, B. Sharp, D. le Sueur, Potential effect of climate change on malaria transmission in Africa. *Lancet* **362**, 1792-1798 (2003).
109. H. E. Tonnang, R. Y. Kangalawe, P. Z. Yanda, Predicting and mapping malaria under climate change scenarios: the potential redistribution of malaria vectors in Africa. *Malar. J.* **9**, 111 (2010).
110. T. K. Yamana, E. A. Eltahir, Projected impacts of climate change on environmental suitability for malaria transmission in West Africa. *Environ. Health Perspect.* **121**, 1179-1186 (2013).
111. S. J. Ryan *et al.*, Mapping Physiological Suitability Limits for Malaria in Africa Under Climate Change. *Vector Borne Zoonotic Dis.* **15**, 718-725 (2015).
112. M. E. Sinka *et al.*, A global map of dominant malaria vectors. *Parasit Vectors* **5**, 69 (2012).

113. T. Lehmann *et al.*, The Rift Valley complex as a barrier to gene flow for Anopheles gambiae in Kenya: the mtDNA perspective. *J. Hered.* **91**, 165-168 (2000).
114. M. A. Slotman *et al.*, Evidence for subdivision within the M molecular form of Anopheles gambiae. *Mol. Ecol.* **16**, 639-649 (2007).
115. E. Oliveira *et al.*, High levels of hybridization between molecular forms of Anopheles gambiae from Guinea Bissau. *J. Med. Entomol.* **45**, 1057-1063 (2008).
116. Y. Lee *et al.*, Ecological and genetic relationships of the Forest-M form among chromosomal and molecular forms of the malaria vector Anopheles gambiae sensu stricto. *Malar. J.* **8**, 75 (2009).
117. C. Ndo *et al.*, Population genetic structure of the malaria vector Anopheles nili in sub-Saharan Africa. *Malar. J.* **9**, 161 (2010).
118. S. Via, Natural selection in action during speciation. *Proc. Natl. Acad. Sci. U. S. A.* **106 Suppl 1**, 9939-9946 (2009).
119. M. Coluzzi, V. Petrarca, M. A. Dideco, Chromosomal Inversion Intergradation and Incipient Speciation in Anopheles-Gambiae. *Boll. Zool.* **52**, 45-63 (1985).
120. T. Lehmann *et al.*, The Rift Valley complex as a barrier to gene flow for Anopheles gambiae in Kenya. *J. Hered.* **90**, 613-621 (1999).
121. L. Kamau, T. Lehmann, W. A. Hawley, A. S. Orago, F. H. Collins, Microgeographic genetic differentiation of Anopheles gambiae mosquitoes from Asembo Bay, western Kenya: a comparison with Kilifi in coastal Kenya. *Am. J. Trop. Med. Hyg.* **58**, 64-69 (1998).
122. E. A. Temu, G. Yan, Microsatellite and mitochondrial genetic differentiation of Anopheles arabiensis (Diptera: Culicidae) from western Kenya, the Great Rift Valley, and coastal Kenya. *Am. J. Trop. Med. Hyg.* **73**, 726-733 (2005).
123. J. M. Drake, J. C. Beier, Ecological niche and potential distribution of Anopheles arabiensis in Africa in 2050. *Malar. J.* **13**, 213 (2014).
124. R. Ramasamy, S. N. Surendran, Possible impact of rising sea levels on vector-borne infectious diseases. *BMC Infect. Dis.* **11**, 18 (2011).
125. B. Caputo *et al.*, Anopheles gambiae complex along The Gambia river, with particular reference to the molecular forms of An. gambiae s.s. *Malar. J.* **7**, 182 (2008).
126. C. Costantini *et al.*, Living at the edge: biogeographic patterns of habitat segregation conform to speciation by niche expansion in Anopheles gambiae. *BMC Ecol.* **9**, 16 (2009).
127. M. Coetzee, H. Cross, Mating experiments with two populations of Anopheles merus Donitz (Diptera: Culicidae). *Journal of Entomological Society of Southern Africa* **46**, 257-259 (1983).
128. A. della Torre *et al.*, Molecular evidence of incipient speciation within Anopheles gambiae s.s. in West Africa. *Insect Mol. Biol.* **10**, 9-18 (2001).
129. G. Favia *et al.*, Molecular identification of sympatric chromosomal forms of Anopheles gambiae and further evidence of their reproductive isolation. *Insect Mol. Biol.* **6**, 377-383 (1997).
130. B. Caputo *et al.*, The "far-west" of Anopheles gambiae molecular forms. *PLoS One* **6**, e16415 (2011).
131. V. A. Mobegi *et al.*, Population genetic structure of Plasmodium falciparum across a region of diverse endemicity in West Africa. *Malar. J.* **11**, 223 (2012).
132. Diego A, Anna U, J. G., Adaptation through chromosomal inversions in Anopheles. *Front Genet* **5**, 129 (2014).

133. Y. T. Toure *et al.*, Ecological genetic studies in the chromosomal form Mopti of Anopheles gambiae s.str. in Mali, west Africa. *Genetica* **94**, 213-223 (1994).
134. T. Lehmann *et al.*, Population Structure of Anopheles gambiae in Africa. *J. Hered.* **94**, 133-147 (2003).
135. S. Blandin, E. A. Levashina, Mosquito immune responses against malaria parasites. *Curr Opin Immunol* **16**, 16-20 (2004).
136. J. A. Scott, W. G. Brogdon, F. H. Collins, Identification of single specimens of the Anopheles gambiae complex by the polymerase chain reaction. *Am. J. Trop. Med. Hyg.* **49**, 520-529 (1993).
137. F. Santolamazza *et al.*, Insertion polymorphisms of SINE200 retrotransposons within speciation islands of Anopheles gambiae molecular forms. *Malar. J.* **7**, 163 (2008).
138. A. L. Smidler, O. Terenzi, J. Soichot, E. A. Levashina, E. Marois, Targeted mutagenesis in the malaria mosquito using TALE nucleases. *PLoS One* **8**, e74511 (2013).
139. T. A. Hall, BioEdit: A user-friendly biological sequence alignment program for Windows 95/98/NT. *Nucleic Acids Symp. Ser.* **40**, 95-98 (1999).
140. H. Kirk, J. R. Freeland, Applications and implications of neutral versus non-neutral markers in molecular ecology. *Int. J. Mol. Sci.* **12**, 3966-3988 (2011).
141. R. D. C. Team. (R Foundation for Statistical Computing, Vienna, Austria, 2016).
142. P. Librado, J. Rozas, DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**, 1451-1452 (2009).
143. J. Rozas, in *Methods Mol. Biol.* (2009), vol. 537, pp. 337-350.
144. M. Clement, D. Posada, K. A. Crandall, TCS: a computer program to estimate gene genealogies. *Mol. Ecol.* **9**, 1657-1659 (2000).
145. D. Posada, Selection of models of DNA evolution with jModelTest. *Methods Mol. Biol.* **537**, 93-112 (2009).
146. J. M. Santorum, D. Darriba, G. L. Taboada, D. Posada, jmodeltest.org: selection of nucleotide substitution models on the cloud. *Bioinformatics* **30**, 1310-1311 (2014).
147. M. Kimura, A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**, 111-120 (1980).
148. K. Tamura, G. Stecher, D. Peterson, A. Filipski, S. Kumar, MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol. Biol. Evol.* **30**, 2725-2729 (2013).
149. S. Guindon *et al.*, New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307-321 (2010).
150. A. J. Drummond, A. Rambaut, BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7**, 214 (2007).
151. B. Korber, in *Computational Analysis of HIV Molecular Sequences,* A. G. Rodrigo, G. H. Learn, Eds. (Kluwer Academic Publishers, Dordrecht, Netherlands, 2000), chap. 4, pp. 55-72.
152. H. DATABASE. SNAP v2.1.1 Synonymous Non-synonymous Analysis Program. https://www.hiv.lanl.gov/content/sequence/SNAP/SNAP.html (2016).
153. L. T. Coggeshall, Anopheles gambiae in Brazil 1930-1940. *Geographical Review* **34**, 308-310 (1944).

154. M. Aboud *et al.*, A genotypically distinct, melanic variant of Anopheles arabiensis in Sudan is associated with arid environments. *Malar. J.* **13**, 492 (2014).
155. S. M. Omer, J. L. Cloudsley-Thompson, Survival of female Anopheles gambiae Giles through a 9-month dry season in Sudan. *Bull. World Health Organ.* **42**, 319-330 (1970).
156. K. A. Crandall, A. R. Templeton, Empirical tests of some predictions from coalescent theory with applications to intraspecific phylogeny reconstruction. *Genetics* **134**, 959-969 (1993).
157. S. Blandin, E. A. Levashina, Thioester-containing proteins and insect immunity. *Mol. Immunol.* **40**, 903-908 (2004).
158. A. Molina-Cruz *et al.*, Some strains of Plasmodium falciparum, a human malaria parasite, evade the complement-like system of Anopheles gambiae mosquitoes. *Proc. Natl. Acad. Sci. U. S. A.* **109**, E1957-1962 (2012).
159. A. Molina-Cruz *et al.*, The human malaria parasite Pfs47 gene mediates evasion of the mosquito immune system. *Science* **340**, 984-987 (2013).
160. U. N. Ramphul, L. S. Garver, A. Molina-Cruz, G. E. Canepa, C. Barillas-Mury, Plasmodium falciparum evades mosquito immunity by disrupting JNK-mediated apoptosis of invaded midgut cells. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 1273-1280 (2015).
161. R. C. Smith, M. Jacobs-Lorena, Malaria parasite Pfs47 disrupts JNK signaling to escape mosquito immunity. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 1250-1251 (2015).
162. M. Manske *et al.*, Analysis of Plasmodium falciparum diversity in natural infections by deep sequencing. *Nature* **487**, 375-379 (2012).
163. B. Franke-Fayard *et al.*, A Plasmodium berghei reference line that constitutively expresses GFP at a high level throughout the complete life cycle. *Mol. Biochem. Parasitol.* **137**, 23-33 (2004).
164. T. Ponnudurai, A. D. Leeuwenberg, J. H. Meuwissen, Chloroquine sensitivity of isolates of Plasmodium falciparum adapted to in vitro culture. *Trop. Geogr. Med.* **33**, 50-54 (1981).
165. V. K. Bhasin, W. Trager, Gametocyte-Forming and Non-Gametocyte-Forming Clones of Plasmodium-Falciparum. *J. Protozool.* **30**, A7-A7 (1983).
166. A. M. Vaughan *et al.*, A transgenic Plasmodium falciparum NF54 strain that expresses GFP-luciferase throughout the parasite life cycle. *Mol. Biochem. Parasitol.* **186**, 143-147 (2012).
167. F. H. Collins *et al.*, Genetic selection of a Plasmodium-refractory strain of the malaria vector Anopheles gambiae. *Science* **234**, 607-610 (1986).
168. C. Harris *et al.*, Polymorphisms in Anopheles gambiae immune genes associated with natural resistance to Plasmodium falciparum. *PLoS Pathog.* **6**, e1001112 (2010).
169. L. Zheng *et al.*, Quantitative trait loci for refractoriness of Anopheles gambiae to Plasmodium cynomolgi B. *Science* **276**, 425-428 (1997).
170. Z. Yang, PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586-1591 (2007).
171. Z. H. Yang, PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**, 555-556 (1997).